# Exploring phonological diversity through principal component analysis

Emily Clem and Lev Michael
*University of California, Berkeley*

eclem@berkeley.edu, levmichael@berkeley.edu

LSA 2017
6 January 2017

# Introduction

- Areal typology is undergoing a revolution as computational methods are being applied to new 'big data' regional datasets (e.g. O'Connor and Muysken, 2014; Reesink et al., 2009; Wichmann and Good, 2014)
- Major goals:
  - Identify typological structure in large regional datasets
  - Identify areal vs. genetic patterns in such datasets (clarify role of language contact in regional typology)

# A tool for areal typology: PCA (correlation and clustering)

- One promising tool to identify large-scale typological patterns is **principal component analysis** (PCA)
- PCA reorganizes a set of correlated variables into new subsets of linearly uncorrelated variables (the 'principal components')
- The principal components (PCs) 'absorb' the correlations in the original variables
    - Examining the PCs informs one about correlations in the data

# A tool for areal typology: PCA (dimension reduction)

- PCs are also **ranked** in terms of what percent of the variance in the dataset each PC accounts for
- In datasets where the original variables exhibit significant correlation:
  - the 'early' PCs (PC1, PC2, ...) account for much of the variance
  - while the later PCs account for little of the variance
- Later PCs can be discarded for many purposes, making PCA a tool for exploratory **dimension reduction** for high-dimension datasets
- Since early PCs account for the major variance in the dataset, they identify major dimensions of **typological differentiation** in a dataset

# PCA and phonological areal typology in South America

- We demonstrate the utility of PCA by applying it to a dataset of South American phonological inventories (SAPhon), to answer the following questions:
  - How is the phonological diversity of South America structured?
    - What are the major typological parameters of differentiation?
  - What areal patterns are detectable?
  - What genetic patterns do we find?

# Preview of results

- We present evidence for a strong areal signal in the Andean and Circum-Andean region and its subregions, separating it from Amazonia
- We will also show evidence for a smaller linguistic area in Northwest Amazonia
- We will argue that languages in South America differ as to whether the locus of phonological contrast is in their consonant system or vowel system
- We will demonstrate that contrasts in nasality and length on vowels are both significant dimensions of differentiation for inventories in South America

# The data: SAPhon

- Our analysis of areality and typological patterning in South American phonologies is based on the SAPhon dataset
- SAPhon (South American Phonological Inventory Database) is an online database of phonological inventories of languages of South America (Michael et al., 2016)
  - `http://linguistics.berkeley.edu/~saphon/en/`
- SAPhon houses phonological inventories for 363 languages
  - All languages for which published inventories are available (plus many for which they aren't): ∼90% of living South American languages
  - 104 more than the number of languages with ISO codes in South America (due to inclusion of extinct languages, and some finer-grained classification)

# Modeling the SAPhon dataset using a vector space

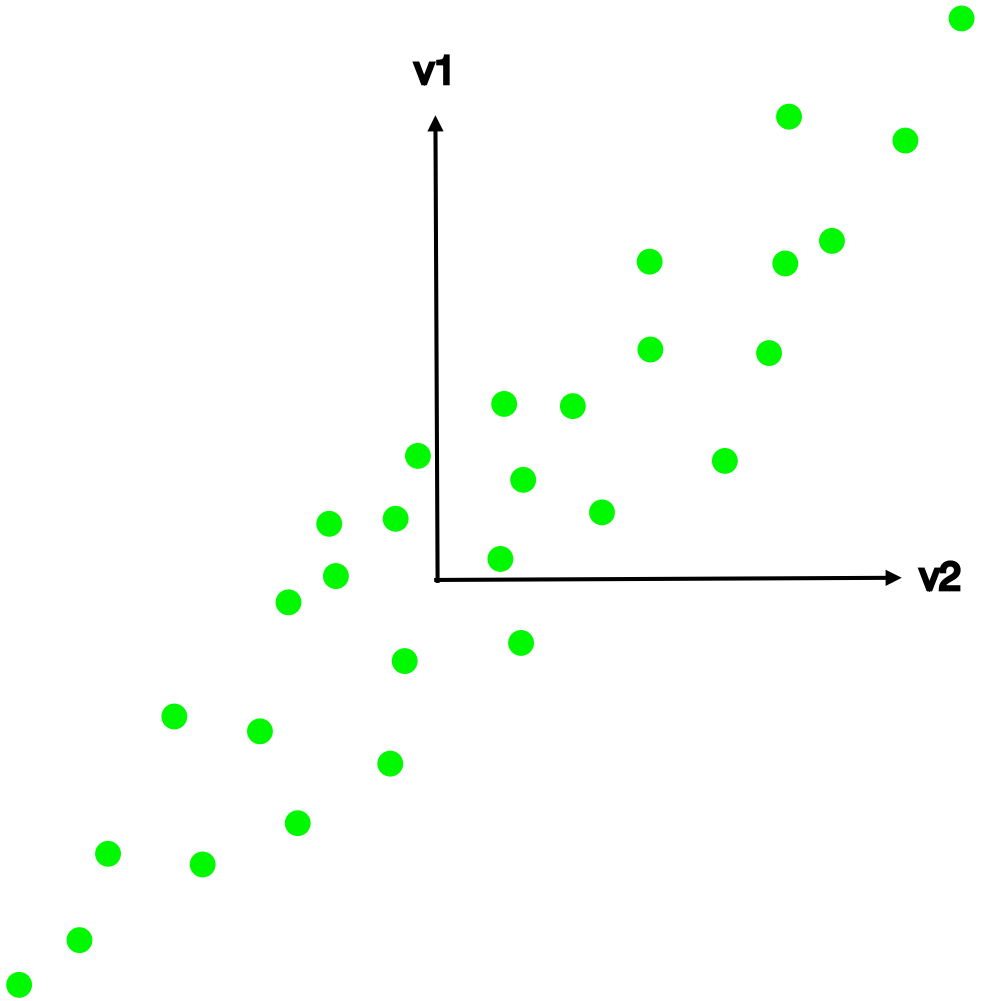
- To apply PCA it is necessary to model the SAPhon inventories as points in a (301-dimensional) vector space:
  - The basis vectors ('axes') that define this space correspond to the 301 segments attested in the SAPhon dataset ($\vec{p}$, $\vec{t}$, $\vec{k}$, $\vec{i}$, $\vec{e}$, $\vec{a}$...)
  - For a given segment vector (e.g., $\vec{p}$), a language exhibits a magnitude of **1** if it has this segment in its inventory, and **0** if it doesn't
  - The combination of these **1** and **0** values for the segment vectors positions each of the 363 SAPhon languages in the vector space

# Understanding PCA

- PCA is a transformation (a rotation) of the original set of basis vectors into a new set of basis vectors (= the **'principal components'**)
  - This rotation eliminates correlations between the basis vectors in the dataset
  - These new vectors are oriented in the 'directions' of greatest variance in the dataset
- As with all rotations in a vector space, the new basis vectors are defined in terms of linear combinations of the old basis vectors

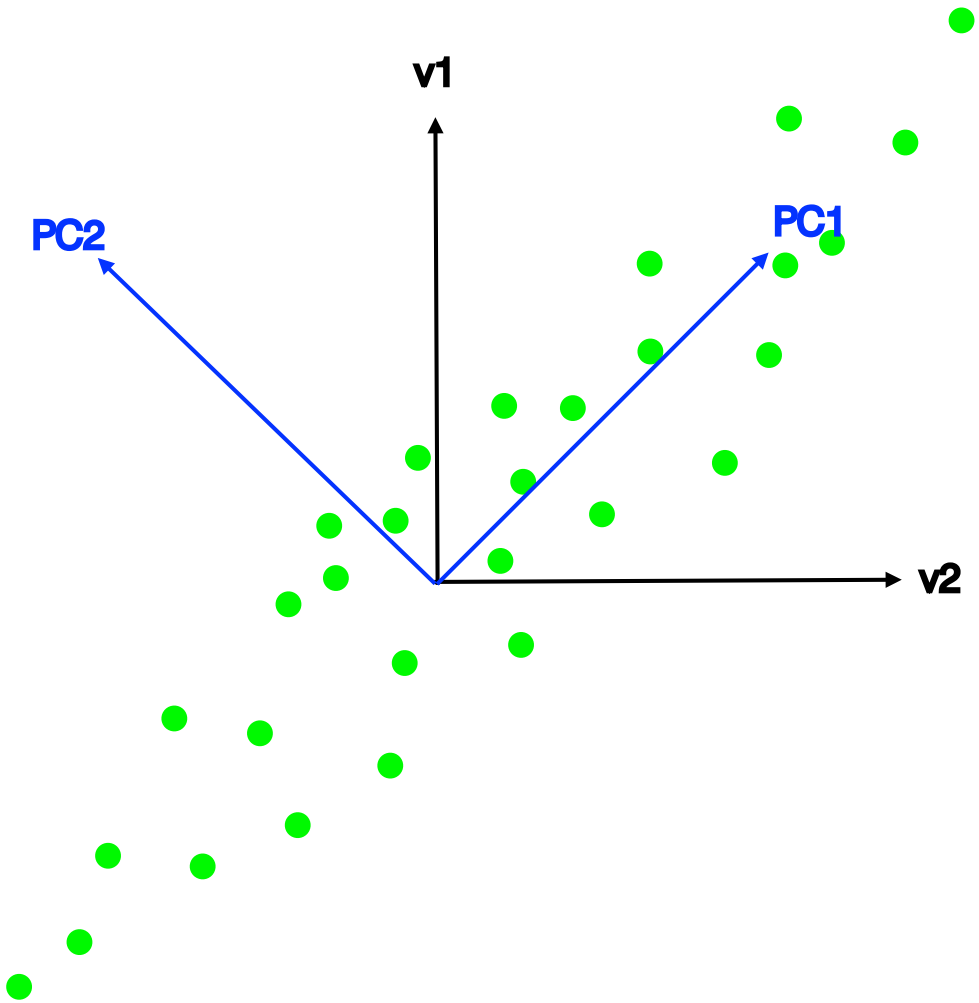
# Understanding PCA

- Dataset with original set of variables (vectors)



# Understanding PCA

- Dataset with original set of variables (vectors) and new PCs
- PC1 oriented in direction of greatest variance in the dataset

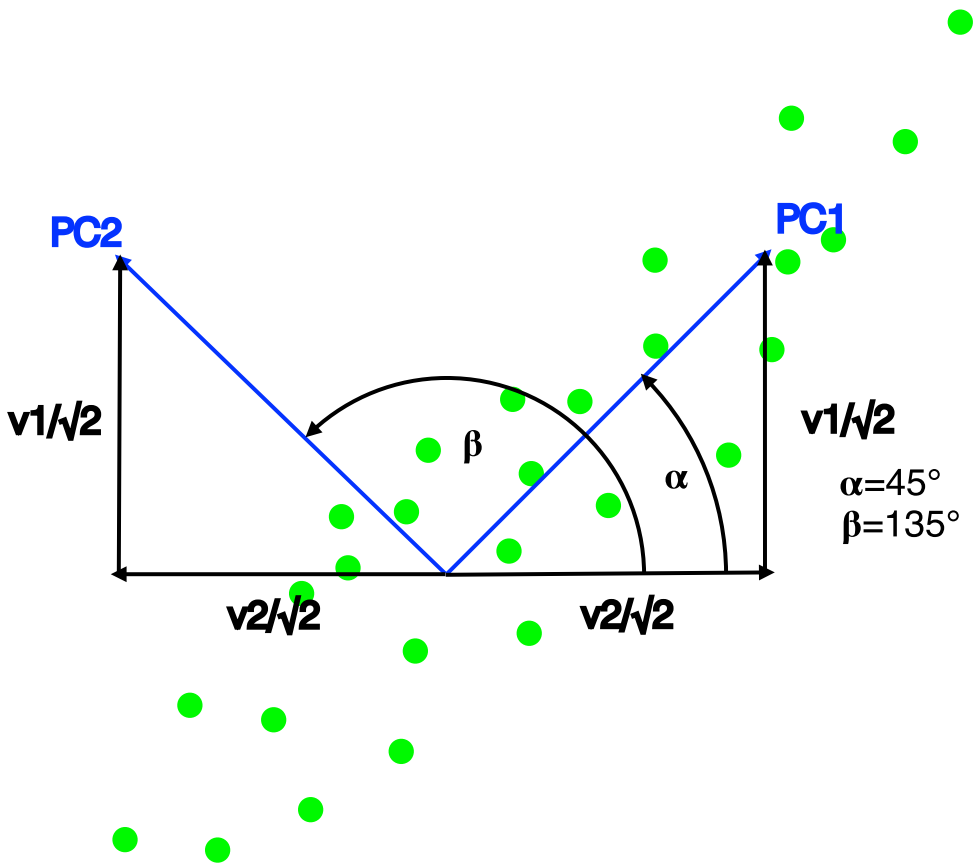


# Understanding PCA

- The PCs are a linear combination of the original basis vectors:

$$(1) \quad \vec{PC1} = sin(\alpha)\vec{v1} + cos(\alpha)\vec{v2} = \frac{1}{\sqrt{2}}\vec{v1} + \frac{1}{\sqrt{2}}\vec{v2}$$

$$(2) \quad \vec{PC2} = sin(\beta)\vec{v1} + cos(\beta)\vec{v2} = \frac{1}{\sqrt{2}}\vec{v1} - \frac{1}{\sqrt{2}}\vec{v2}$$

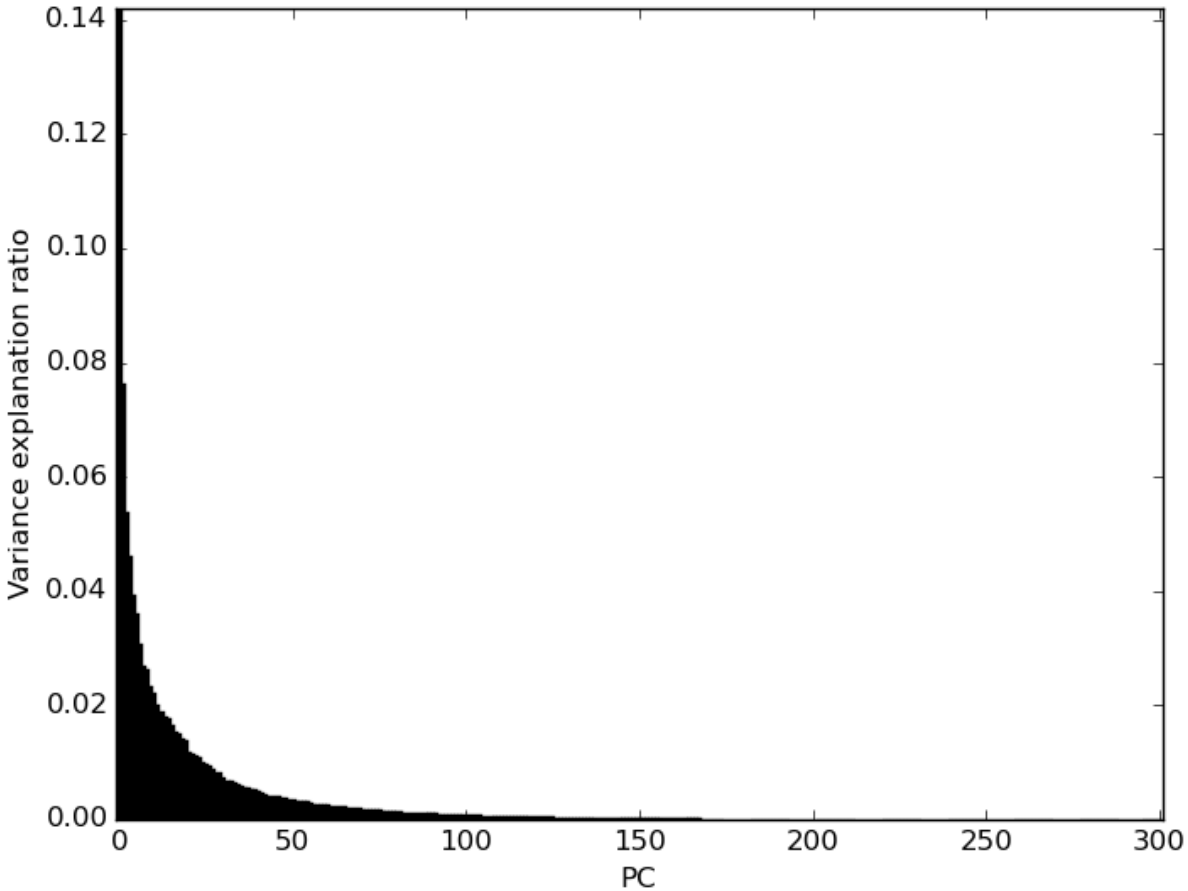


# SAPhon PCs

- The PCs obtained by carrying out PCA on the original SAPhon basis vectors are thus linear combinations of the segment vectors
- The linear sum specifies the positive and negative weights accorded each segment in calculating the given PC:
  - e.g. PC1 $= 0.140$**l** $+ 0.137$**ts** $+ 0.128$**ʃ** $+ 0.114$**ʌ** $+ 0.122$**aː** $\ldots - 0.359$**ĩ** $- 0.348$**ã** $- 0.315$**ẽ** $- 0.312$**ũ** $- 0.306$**õ**
- Each PC resembles a pair of weighted phonological inventories:
  - A **positive inventory** that characterizes the positive extremum of that PC
  - A **negative inventory** that characterizes the negative extremum of that PC

# Variance explained by PCs

- A comparatively small number of the 301 PCs are responsible for explaining the majority of the variance in the dataset
- PC1 = 14.2%, PC2 = 7.6%, PC3 ≈ 5.4%, ...

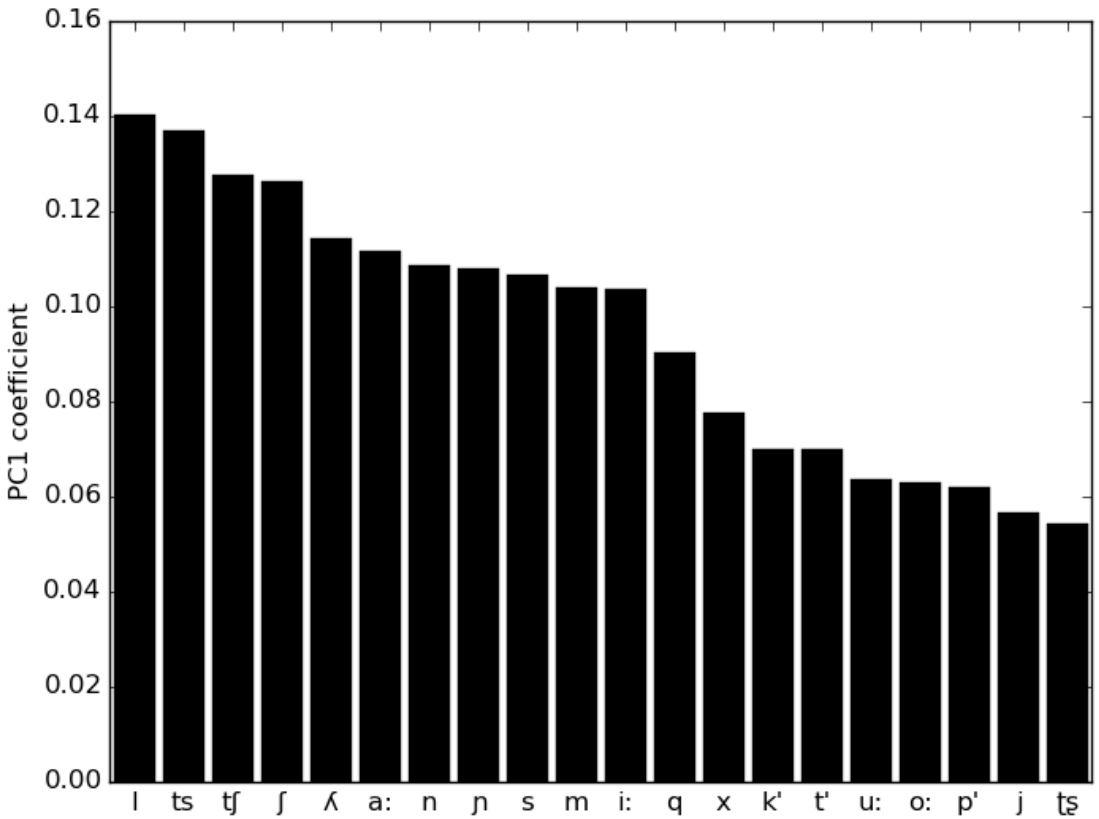


# Variance explained by PCs

- A comparatively small number of the 301 PCs are responsible for explaining the majority of the variance in the dataset
- PC1 = 14.2%, PC2 = 7.6%, PC3 $\approx$ 5.4%, ...
- Due to the rapid decrease in variance explained by the successive PCs, we can focus on the largest PCs to identify the significant 'typological structure' of the SAPhon phonological inventories
- For each of the first 5 PCs, we can examine both areal and genetic patterns that are revealed

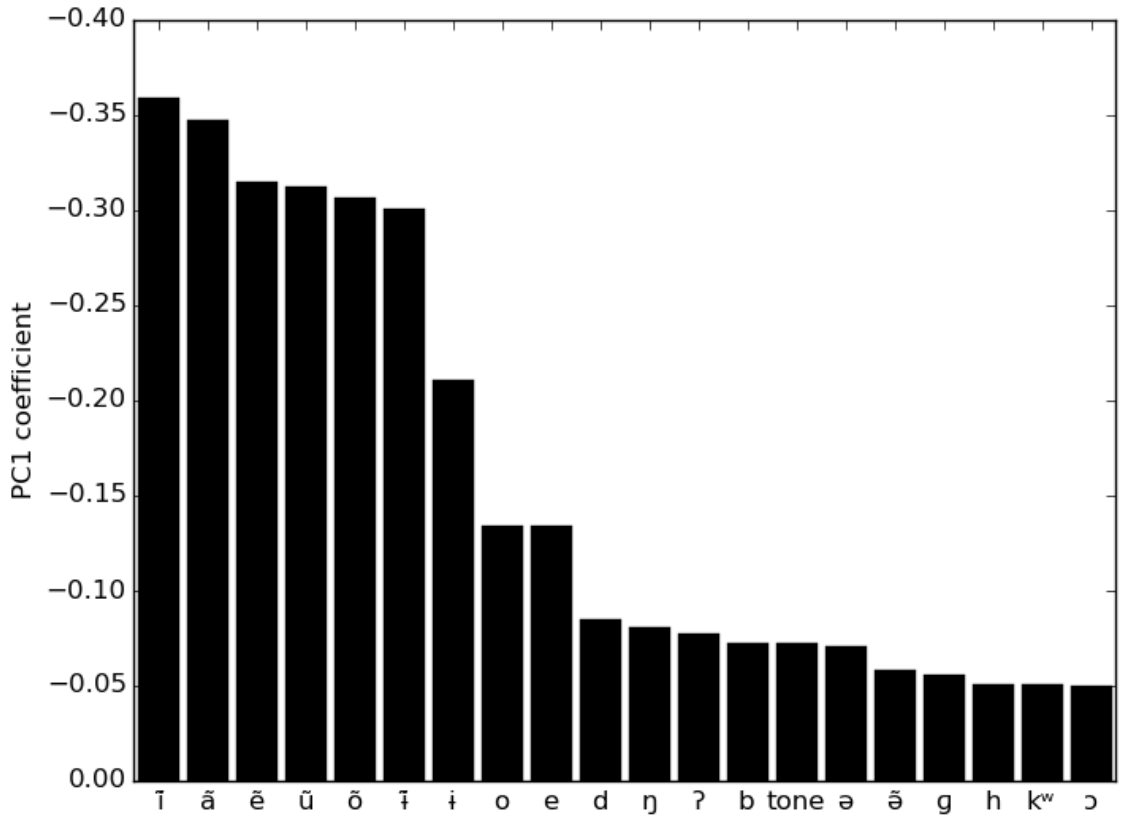
# PC1: positive coefficients

- PC1 explains 14.2% of the variance in the data
- The segments with the largest positive coefficients include affricates, palatals, and laterals



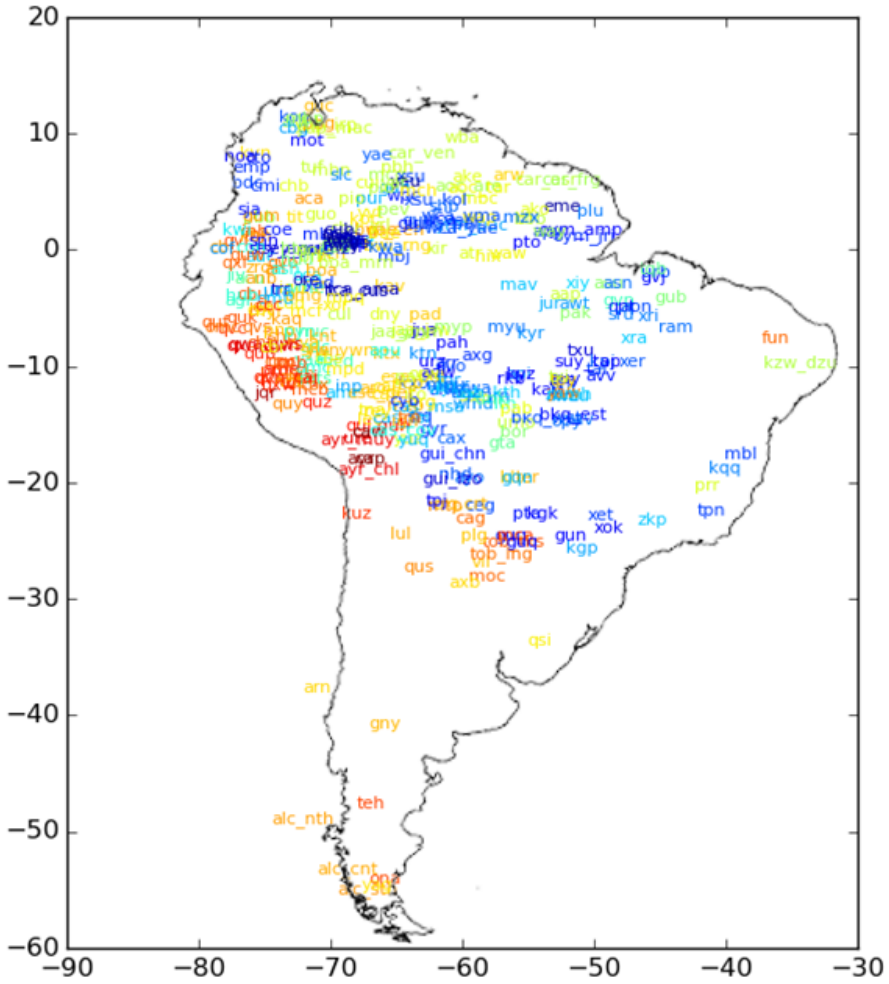
# PC1: negative coefficients

- Negative coefficients are slightly larger than positive coefficients
- The segments with the largest negative coefficients include nasal vowels, ɨ, and mid vowels



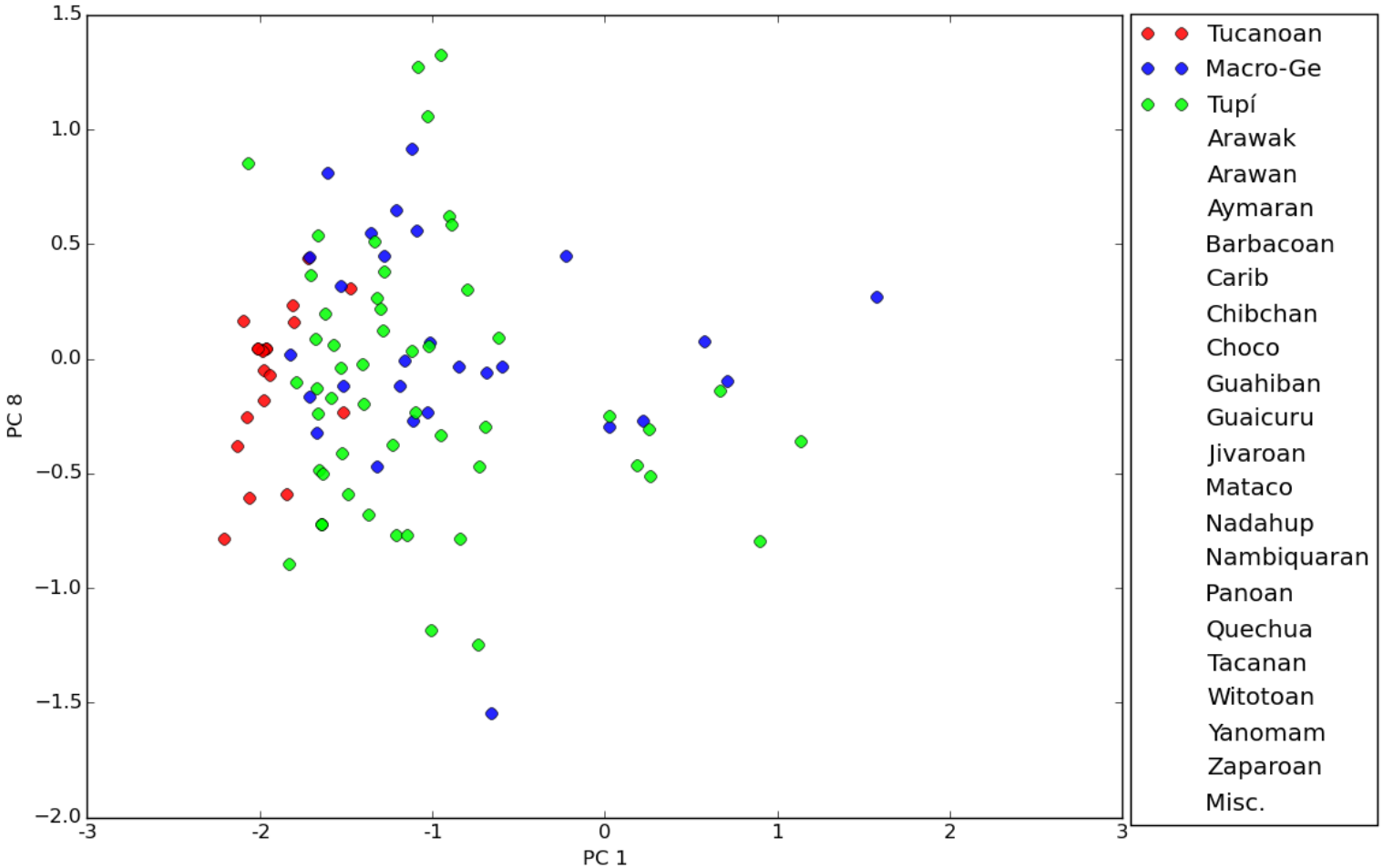
# PC1: areal signal

- PC1 yields a strong positive signal in the Andean and Circum-Andean region



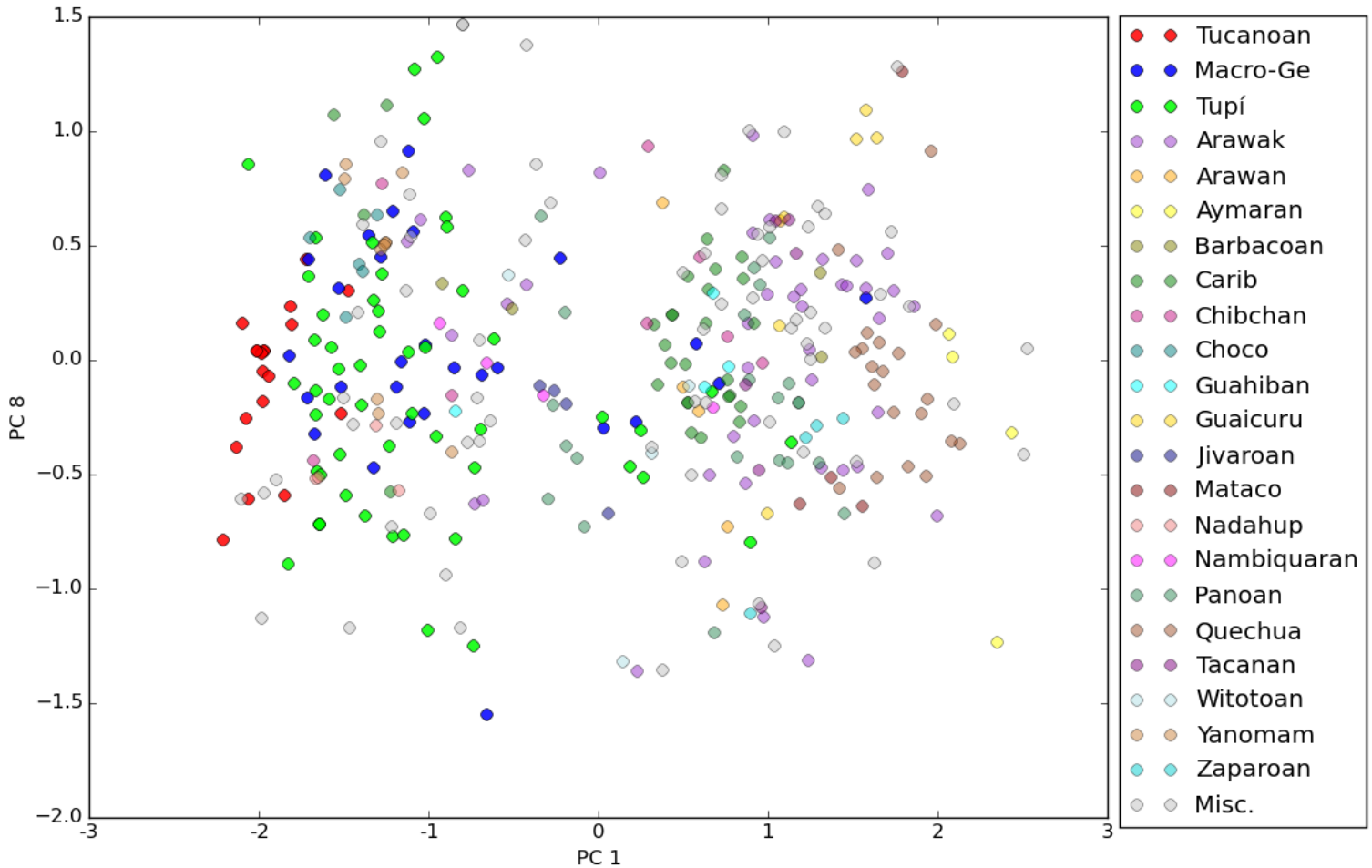
# PC1: genetic signal

- PC1 shows a negative genetic signal including the Tucanoan ($p<1.0E-13$), Tupí ($p<1.0E-11$), and Macro-Ge ($p<1.0E-4$) families (using Kolmogorov-Smirnov Test)



# PC1: genetic signal

- PC1 shows a negative genetic signal including the Tucanoan ($p<1.0E-13$), Tupí ($p<1.0E-11$), and Macro-Ge ($p<1.0E-4$) families (using Kolmogorov-Smirnov Test)

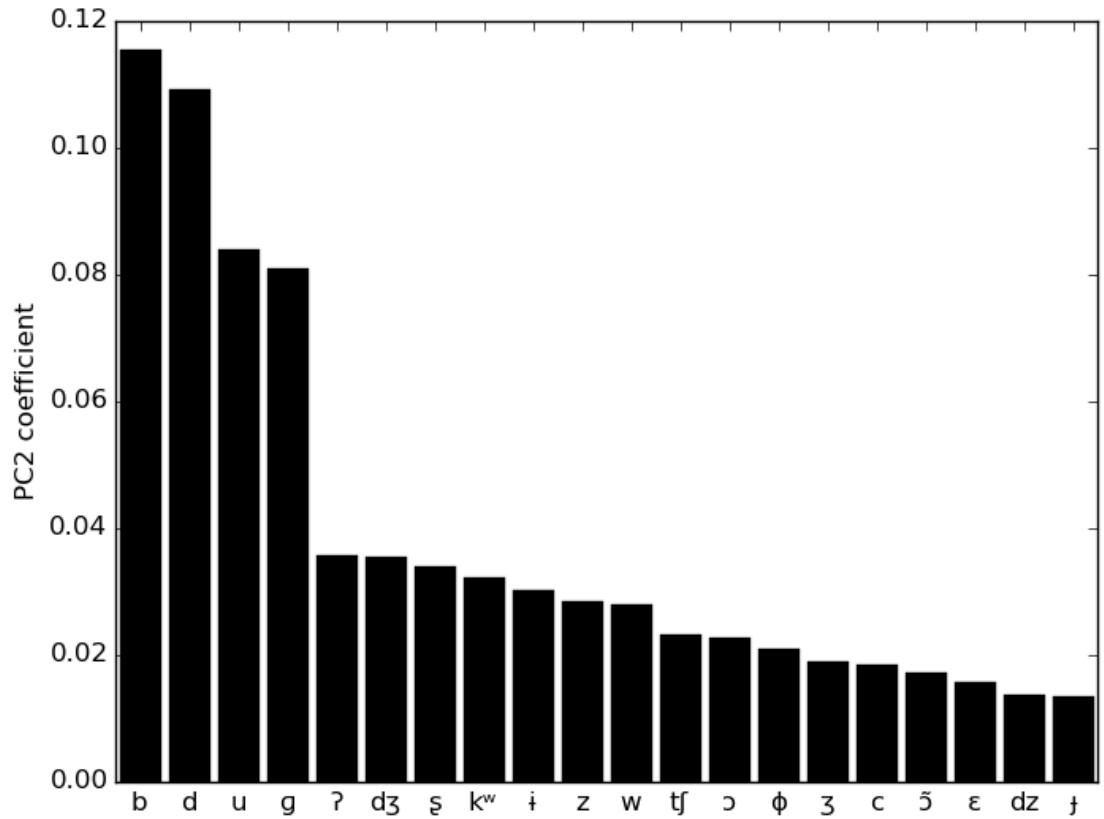


# PC1: summary

- Positive segments: alveolar and palatal laterals, affricates, fricatives, and nasals
- Negative segments: nasal vowels, ɨ
- Positive component yields a strong areal signal in the Andes and Circum-Andean area, including Patagonia
  - Independently identifies this phonological area, first found using a Naive Bayesian Classifier method (Michael et al., 2014)
- Negative component shows a genetic signal from Tucanoan, Macro-Ge, and Tupí families

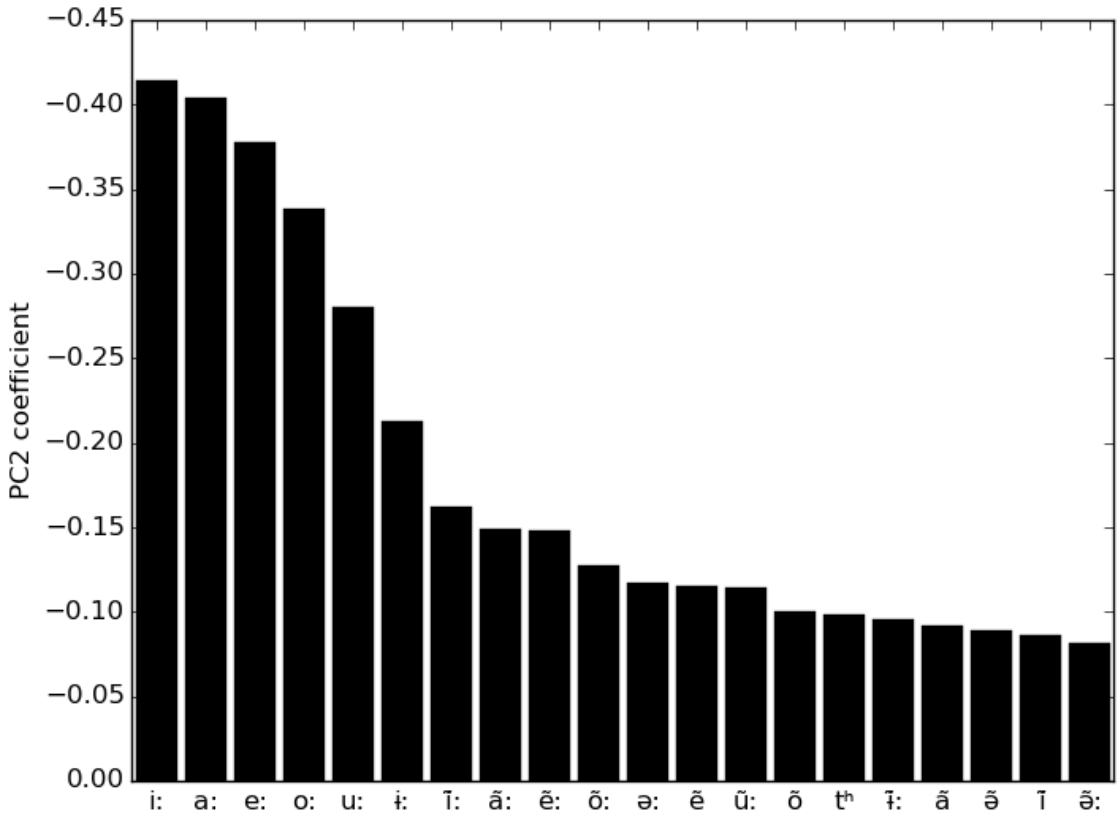
# PC2: positive coefficients

- PC2 explains 7.6% of the variance in the data
- The segments with the largest positive coefficients include the voiced stops



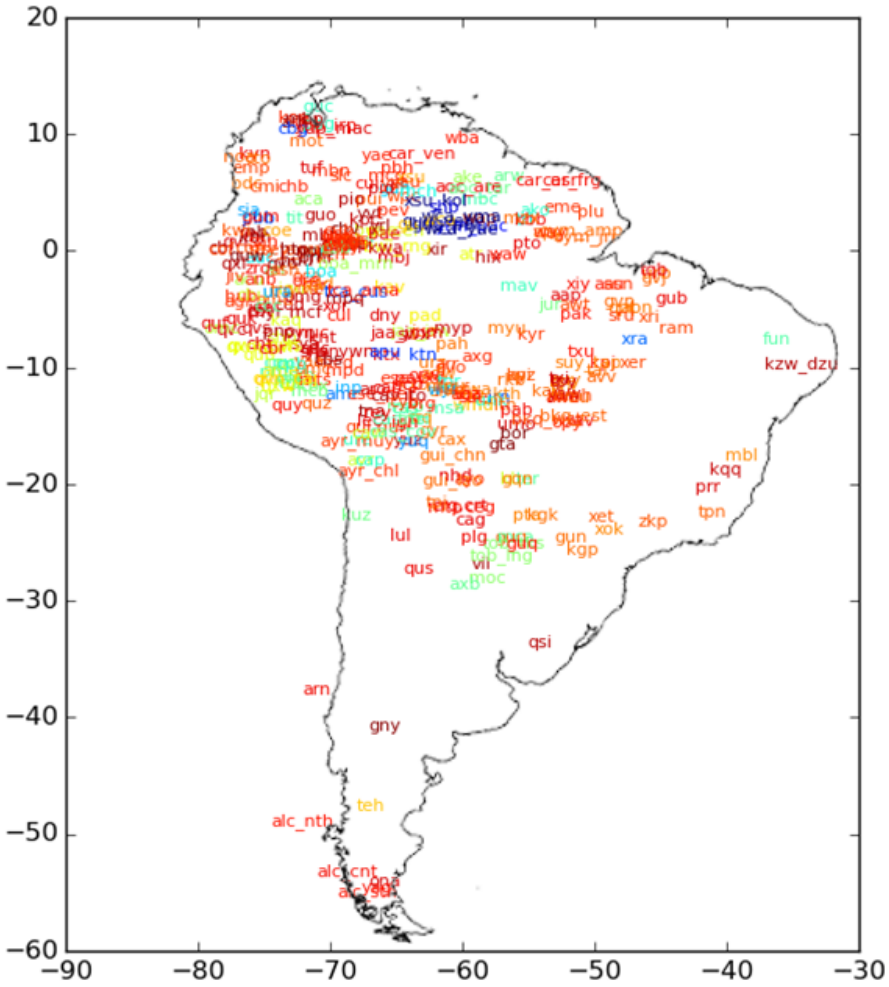
# PC2: negative coefficients

- Negative coefficients are larger than positive coefficients
- All of the segments with the largest negative coefficients are long vowels



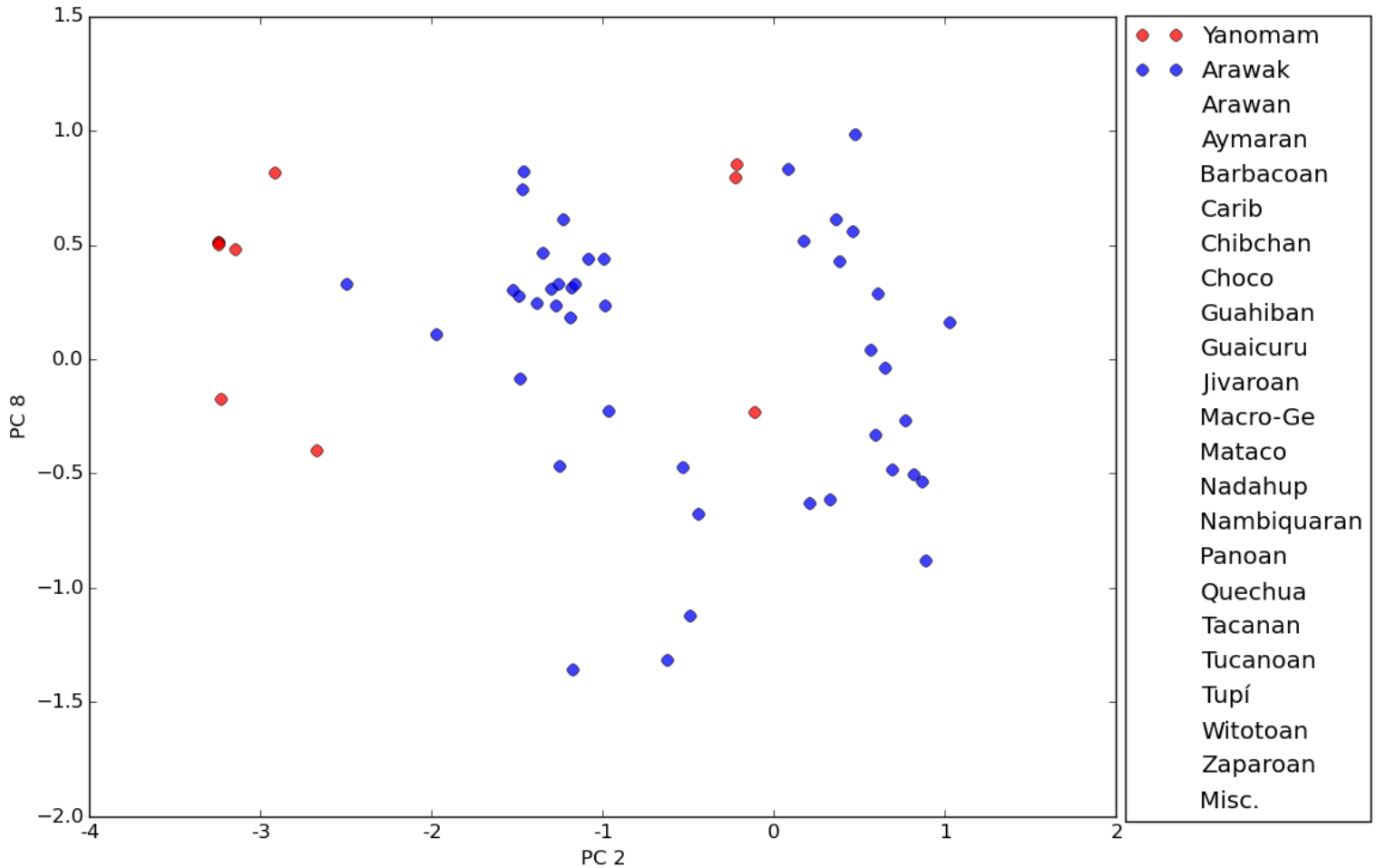
# PC2: areal signal

- PC2 does not display a strong areal signal distinct from genetic relationships



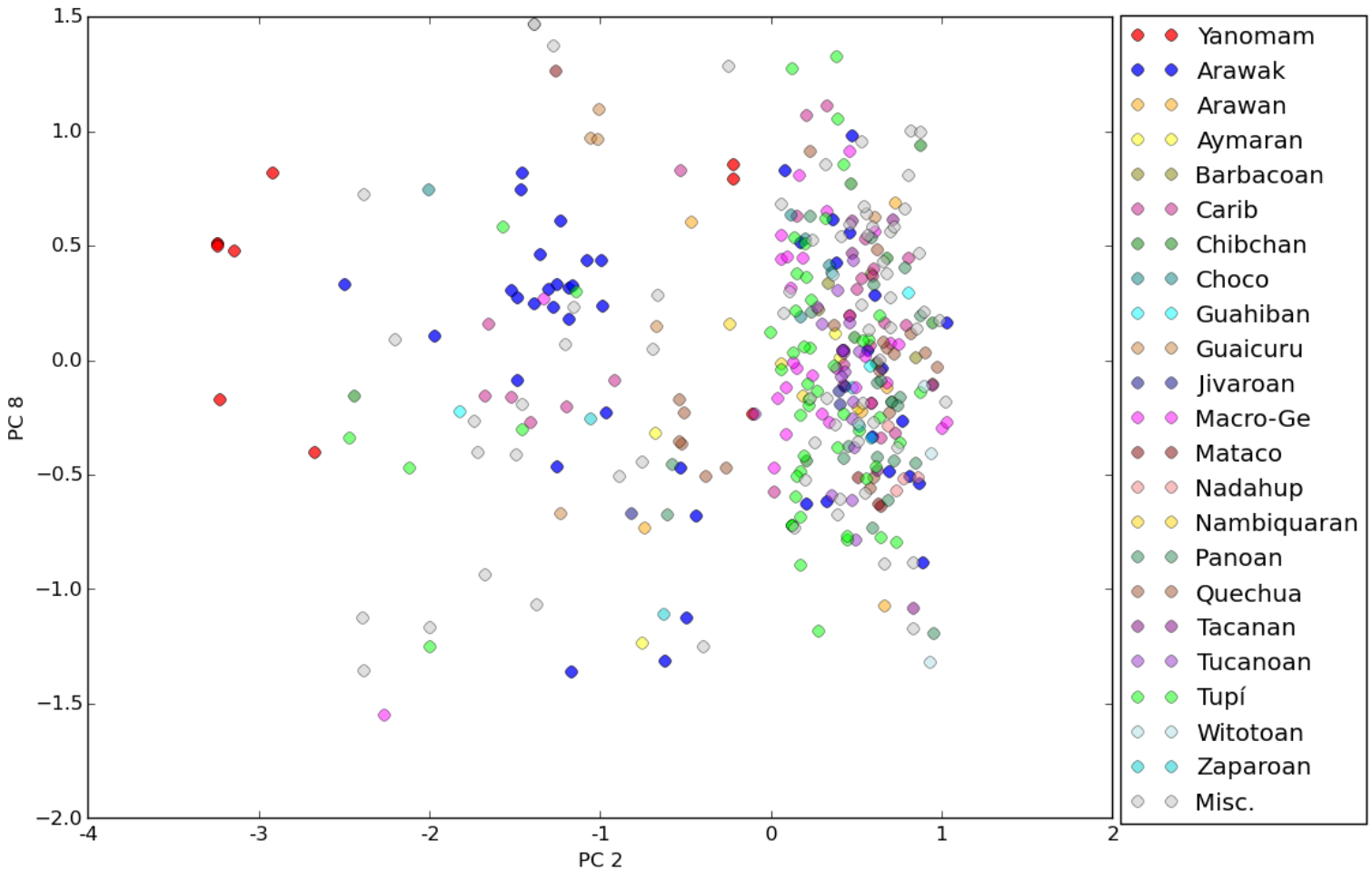
# PC2: genetic signal

- PC2 shows negative genetic signal from Arawak ($p<1.0E-5$) and Yanomam ($p<1.0E-4$)



# PC2: genetic signal

- PC2 shows negative genetic signal from Arawak ($p<1.0E-5$) and Yanomam ($p<1.0E-4$)

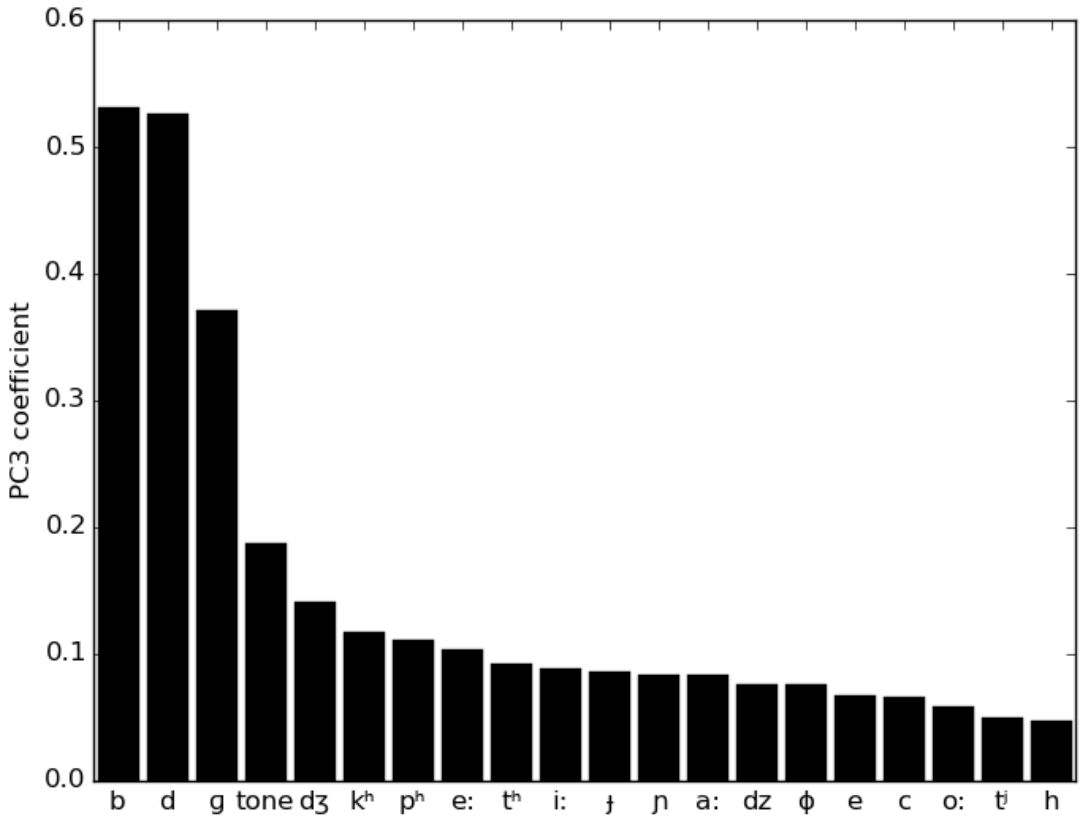


# PC2: summary

- Positive segments: voiced stops
- Negative segments: long vowels
- Negative component shows a strong genetic signal associated with Yanomam, and other families also cluster together
- We see a large negative dispersion with the most strongly negative languages displaying vowel length contrasts for many vowels

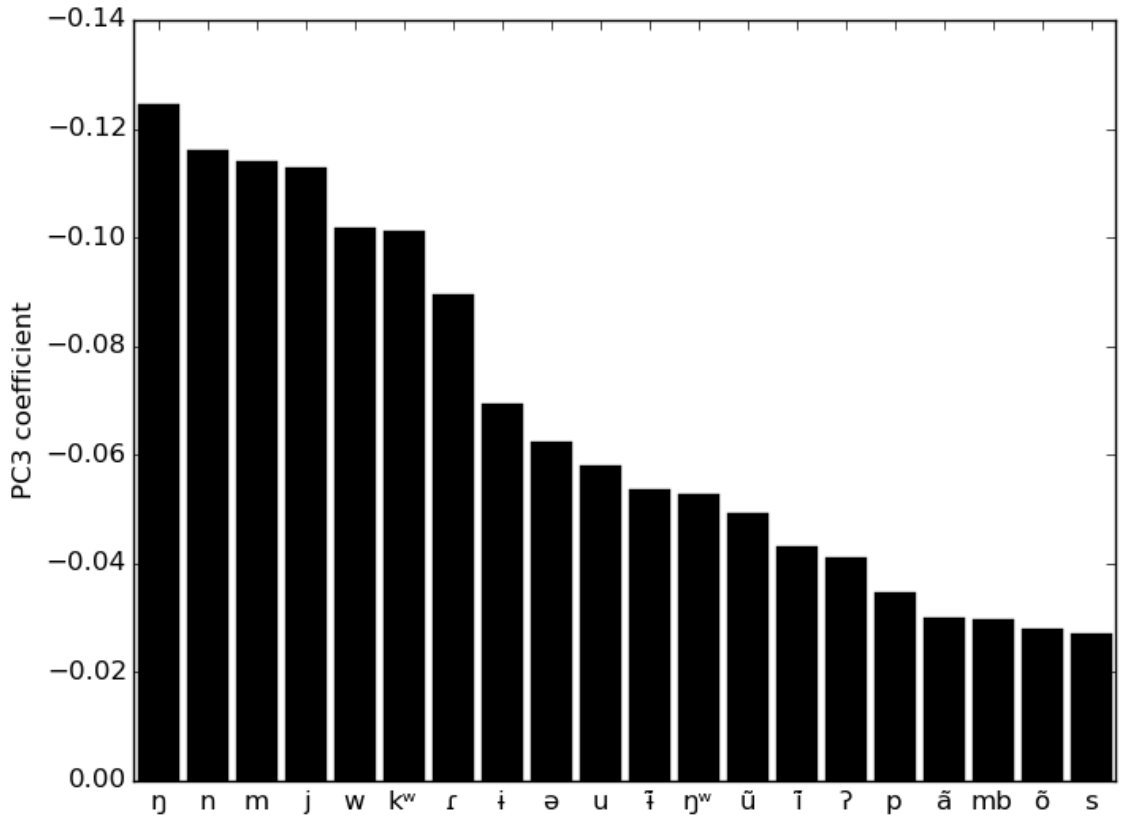
# PC3: positive coefficients

- PC3 explains 5.4% of the variance in the data
- The segments with the largest positive coefficients include the voiced stops, tone, and aspirated stops
- Positive coefficients are much larger than negative coefficients



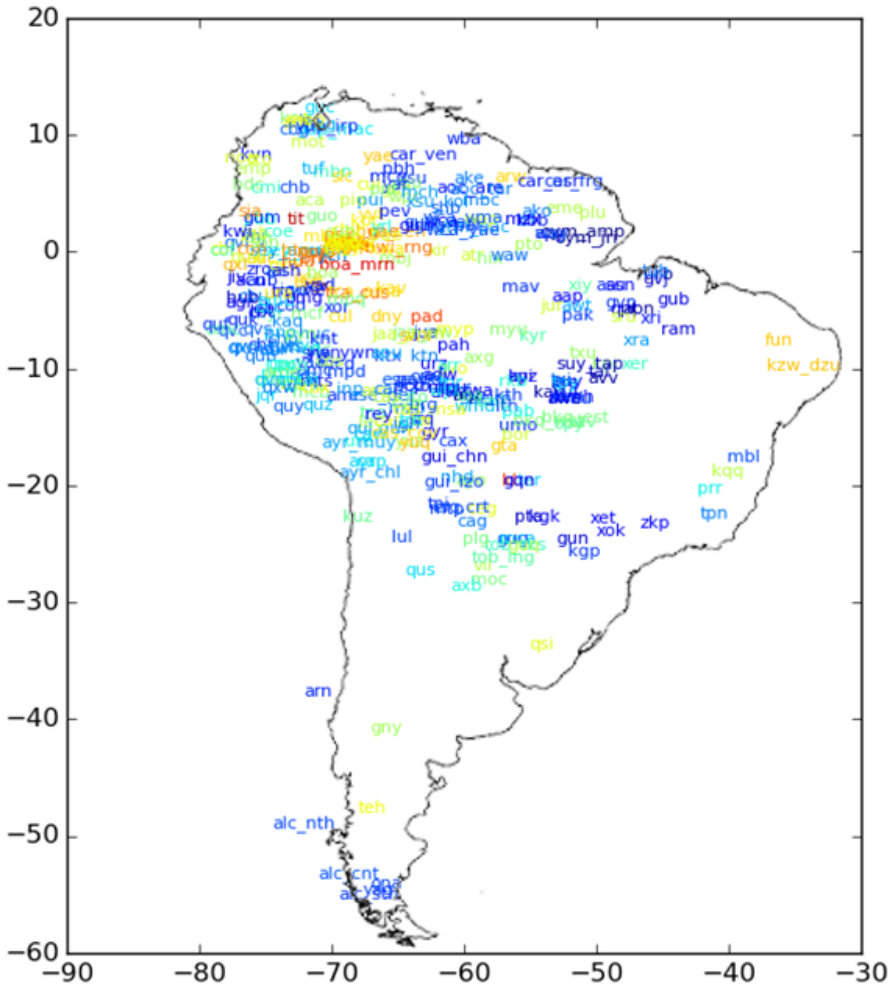
# PC3: negative coefficients

- The segments with the largest negative coefficients include the nasal stops and approximants



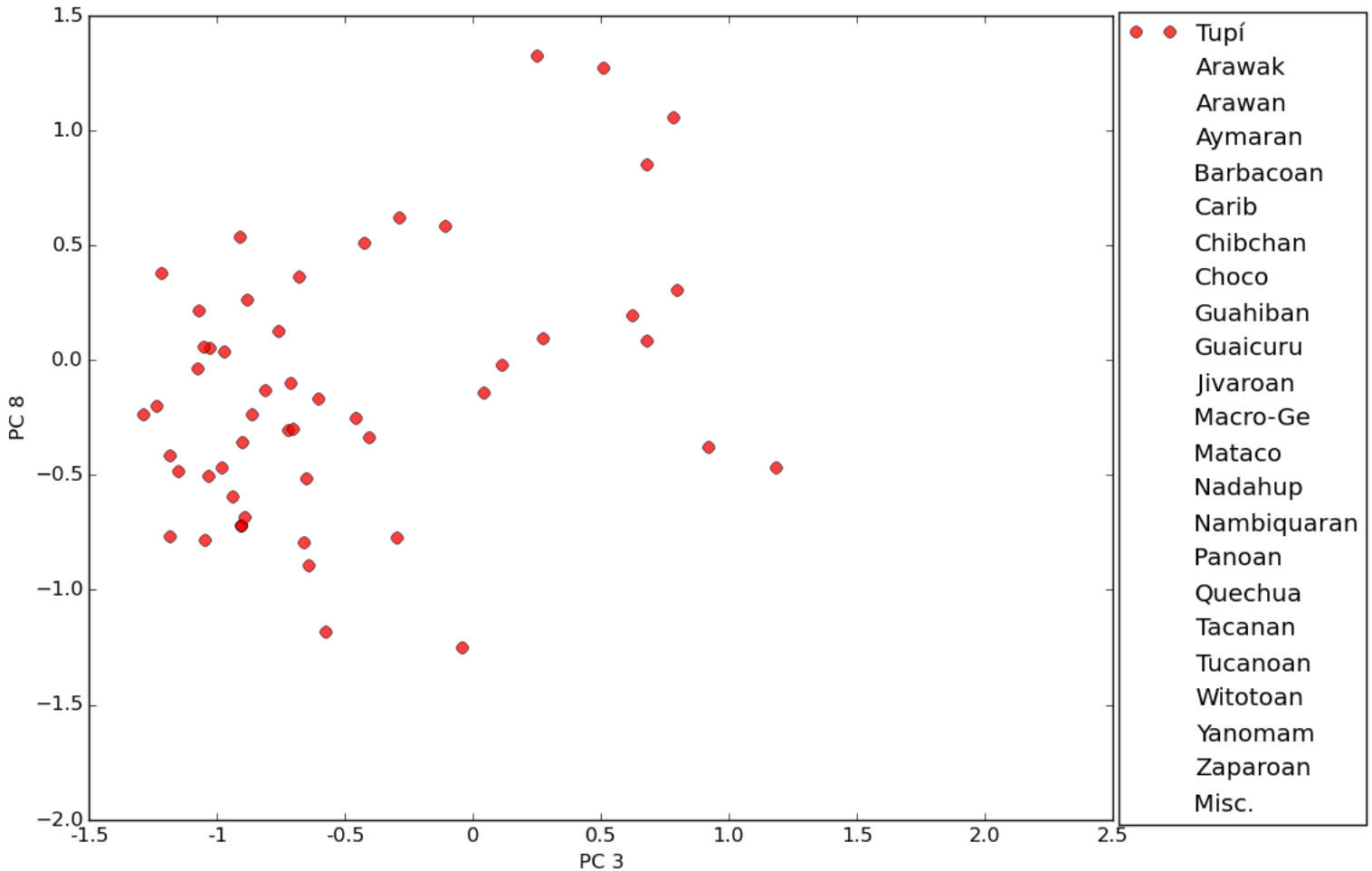
# PC3: areal signal

- PC3 shows a strong positive signal in Northwest Amazonia



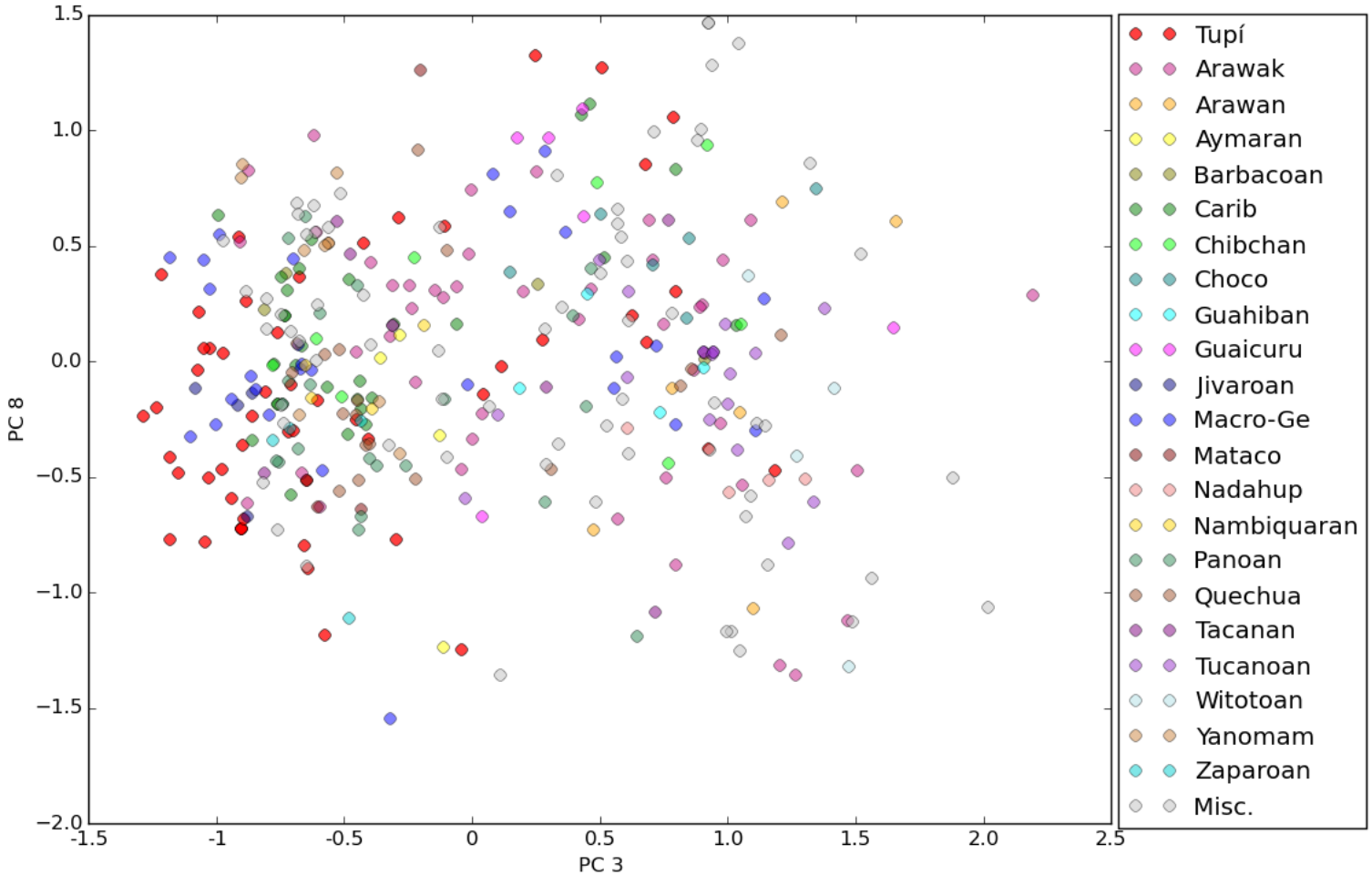
# PC3: genetic signal

- PC3 shows a strong negative genetic signal associated with Tupí ($p<1.0E-5$)



# PC3: genetic signal

- PC3 shows a strong negative genetic signal associated with Tupí ($p<1.0E-5$)

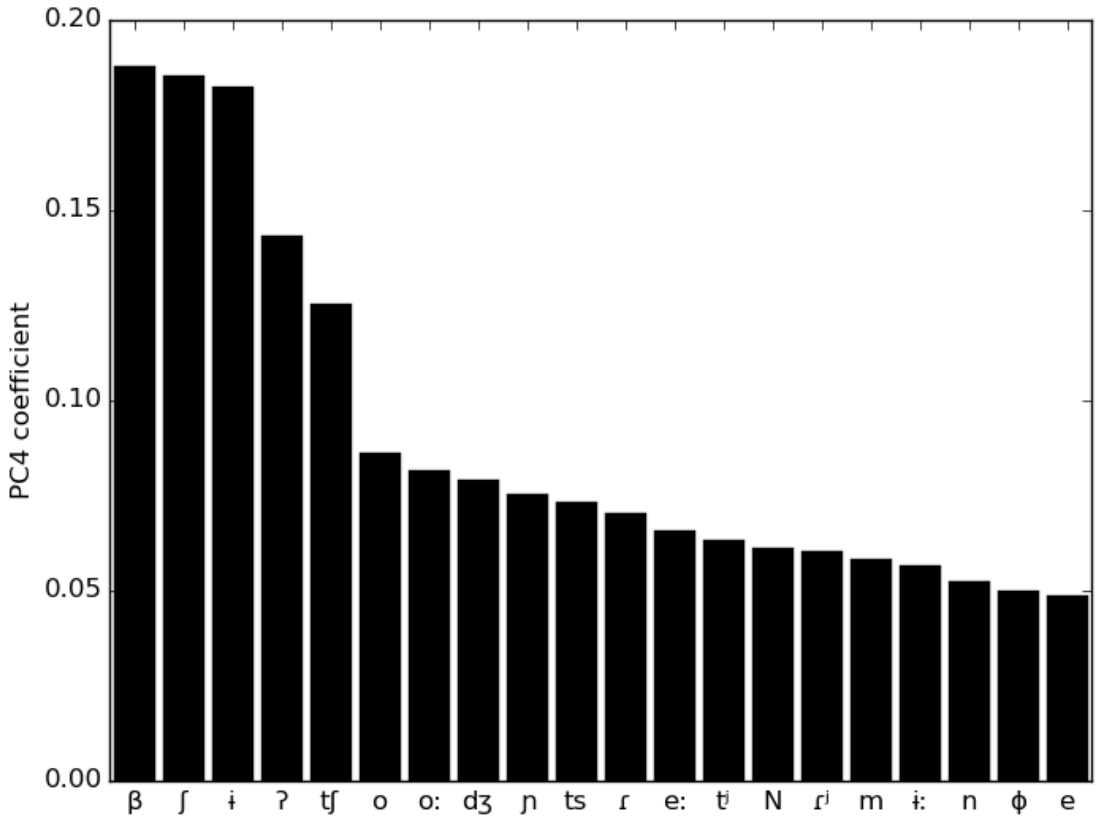


# PC3: summary

- Positive segments: voiced stops, tone
- Negative segments: nasal stops, approximants
- Positive component yields a strong areal signal in Northwest Amazonia
  - Identifies this well known linguistic area (see, e.g. Aikhenvald, 2002) on the basis of phonological inventories alone
- This positive signal reflects languages that have processes of nasal harmony rather than underlying nasal stops
- Negative component shows a genetic signal from Tupí

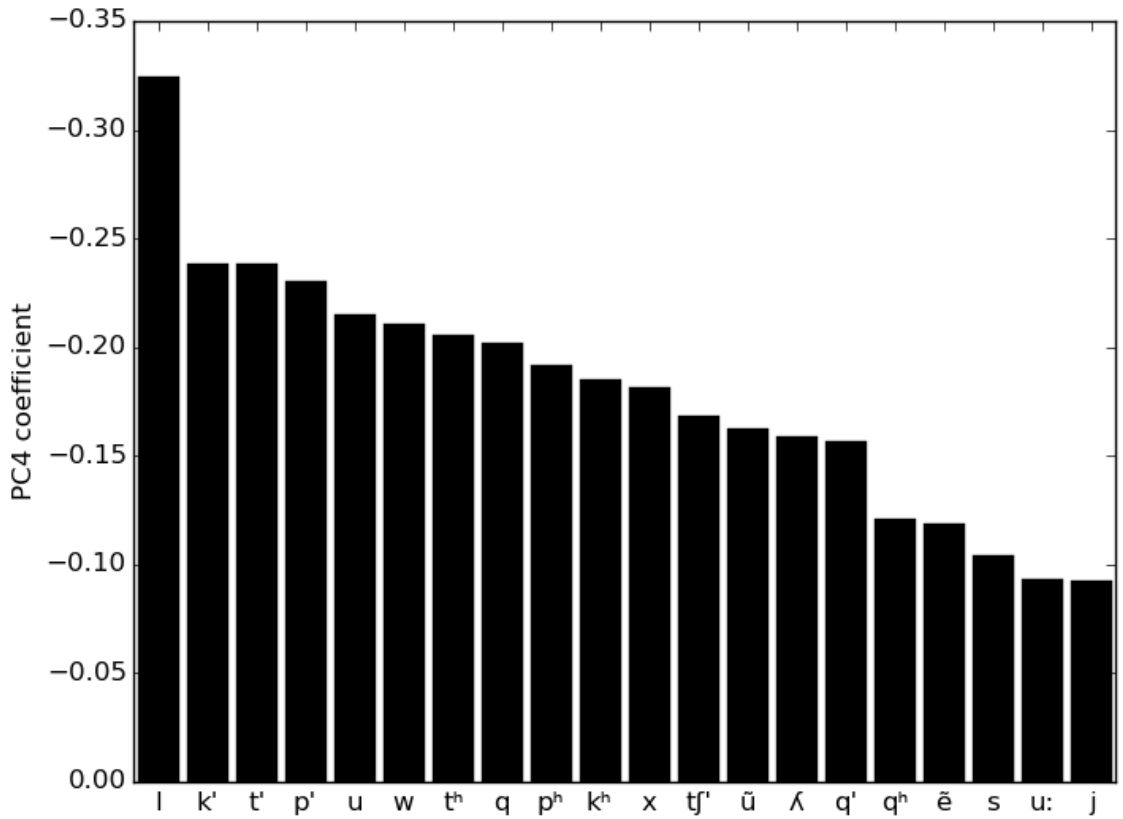
# PC4: positive coefficients

- PC4 explains 4.6% of the variance in the data
- The segments with the largest positive coefficients include β, ɨ, palatals, and mid vowels



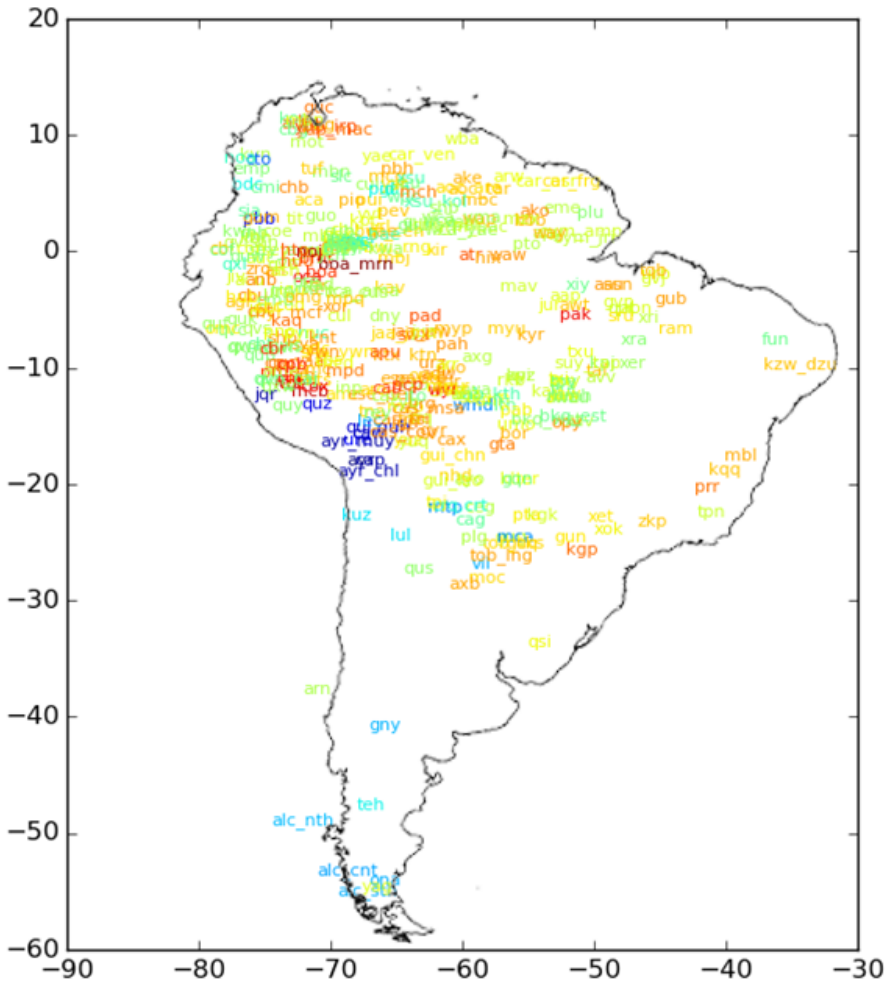
# PC4: negative coefficients

- Negative coefficients are larger than positive coefficients
- The segments with the largest negative coefficients are ejectives, aspirated stops, laterals, and uvulars



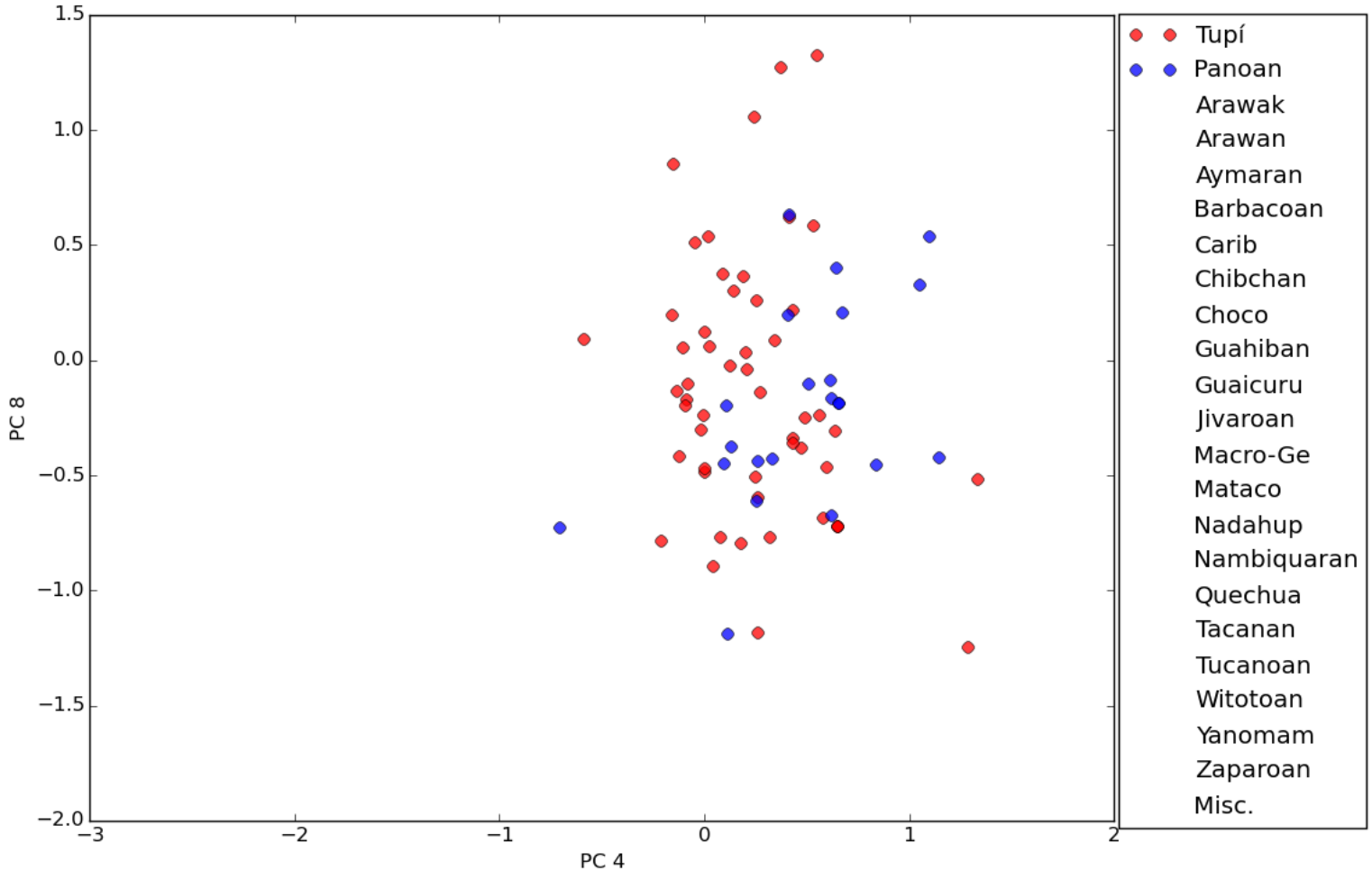
# PC4: areal signal

- PC4 shows a strong negative signal in the Southern Andean and Circum-Andean region, including Patagonia



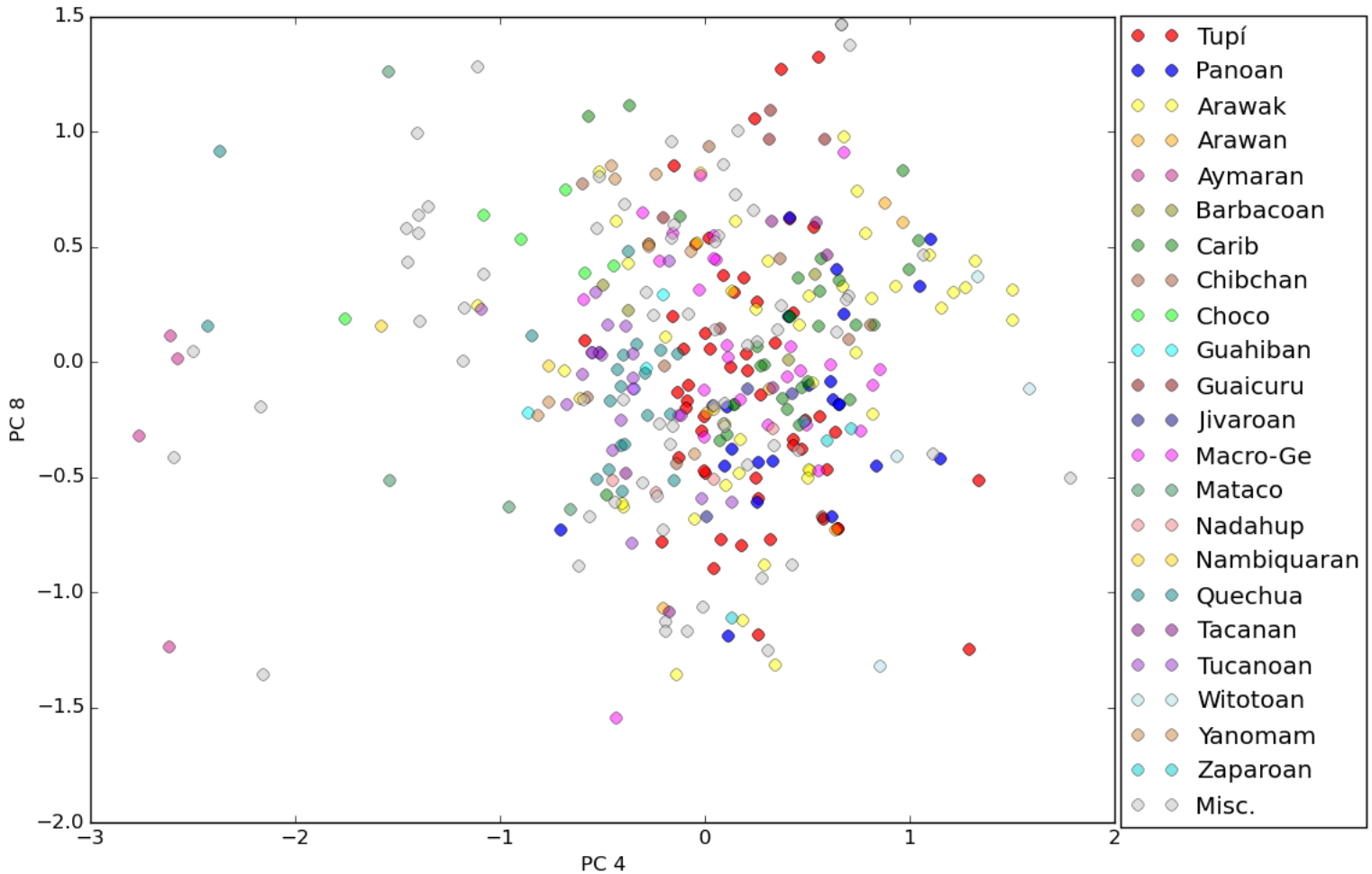
# PC4: genetic signal

- PC4 shows a positive genetic signal associated with Tupí ($p<1.0E-5$) and Panoan ($p<1.0E-4$)



# PC4: genetic signal

- PC4 shows a positive genetic signal associated with Tupí ($p<1.0E-5$) and Panoan ($p<1.0E-4$)

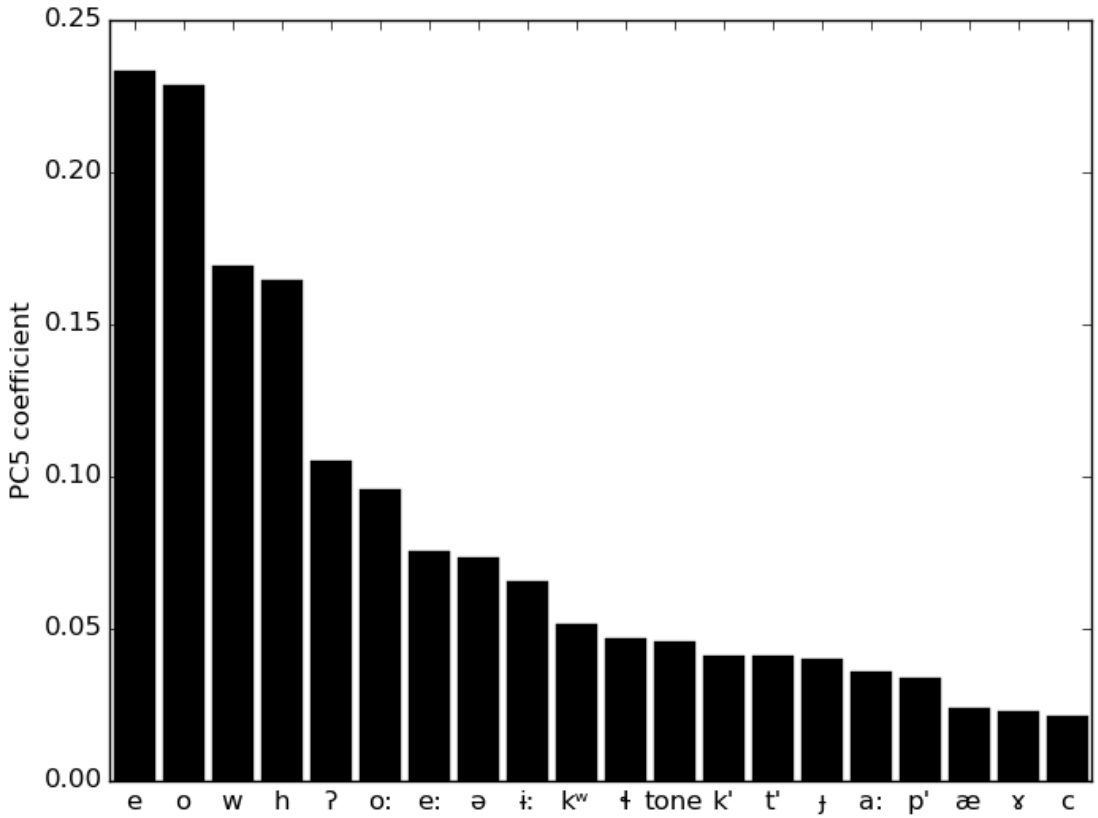


# PC4: summary

- Positive segments: β, ɨ, palatals
- Negative segments: ejectives, aspirated stops, laterals, uvulars
- Negative component shows further support for a strong areal signal in Southern Andean and Circum-Andean region
  - Identifies this important sub-area of the Andean and Circum-Andean area also found by Michael et al. (2014)
- Positive component shows a genetic signal from Tupí and Panoan

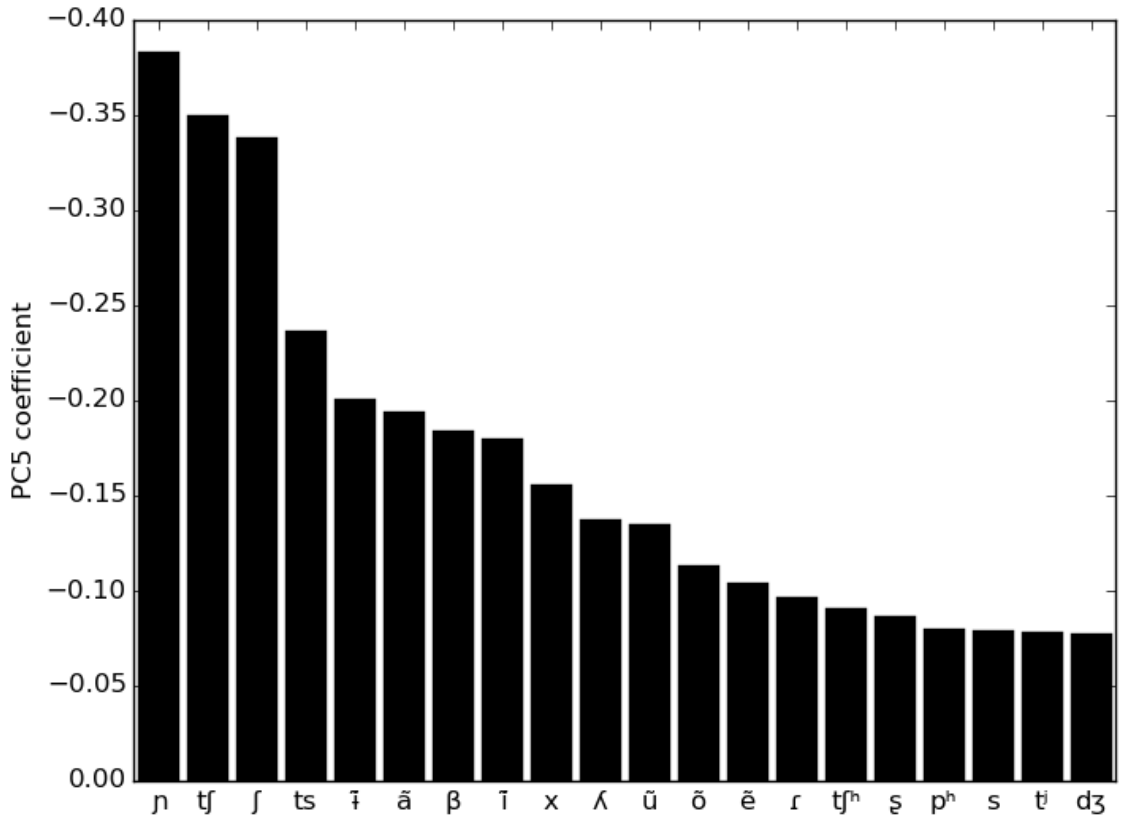
# PC5: positive coefficients

- PC5 explains 3.9% of the variance in the data
- The segments with the largest positive coefficients are mid vowels, w, and glottals



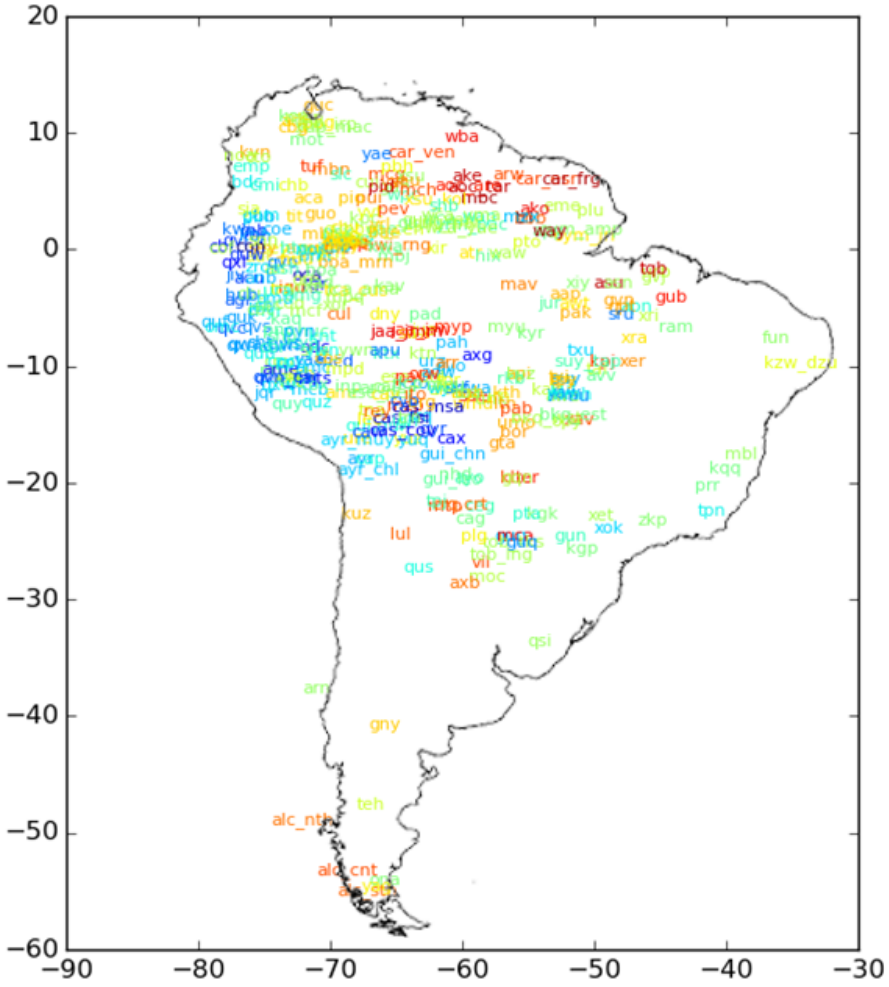
# PC5: negative coefficients

- Negative coefficients are slightly larger than positive coefficients
- The segments with the largest negative cofficients are palatals, affricates, and ɨ



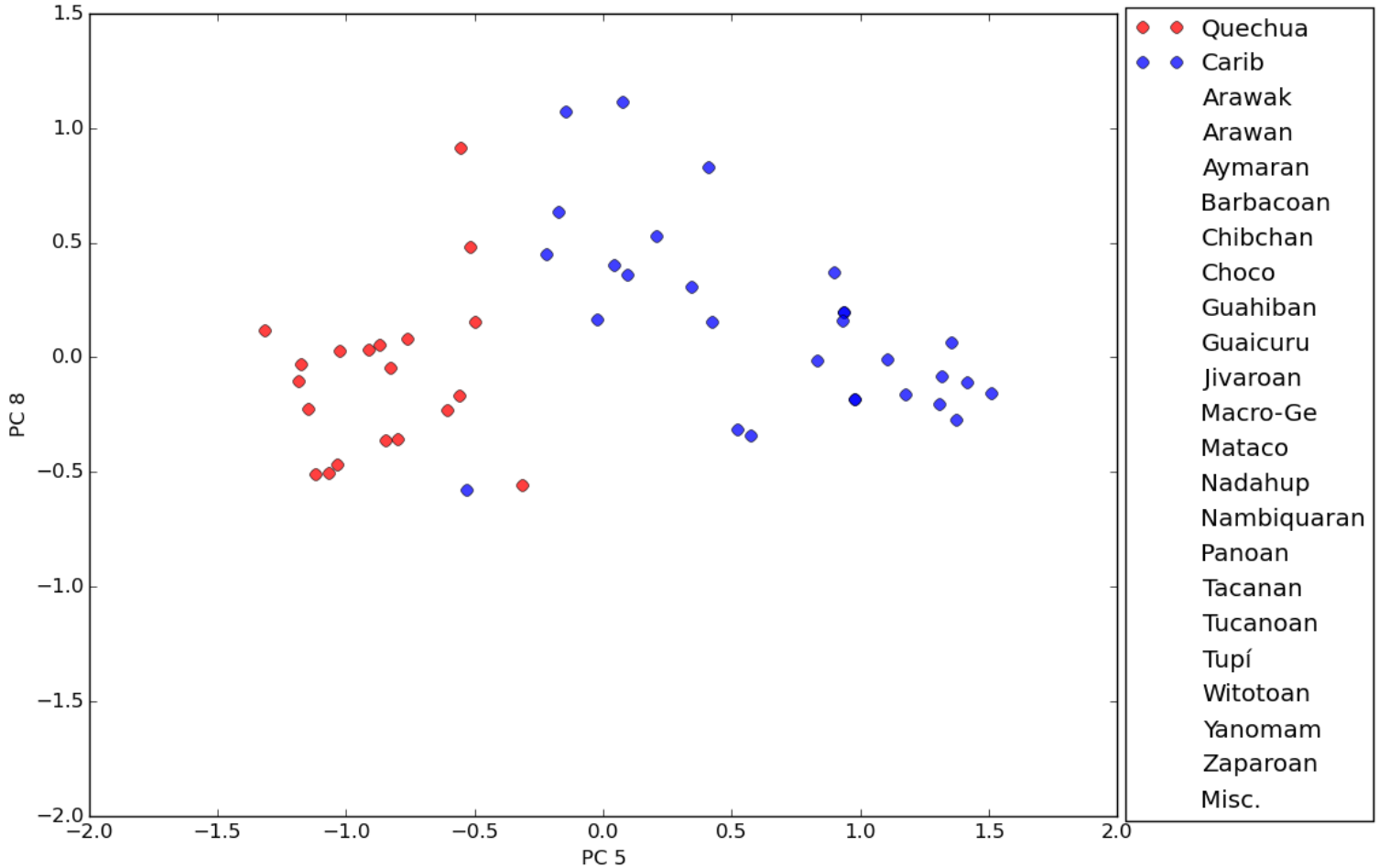
# PC5: areal signal

- PC5 shows a negative signal in the Central-Northern Andean region and a positive signal in the east



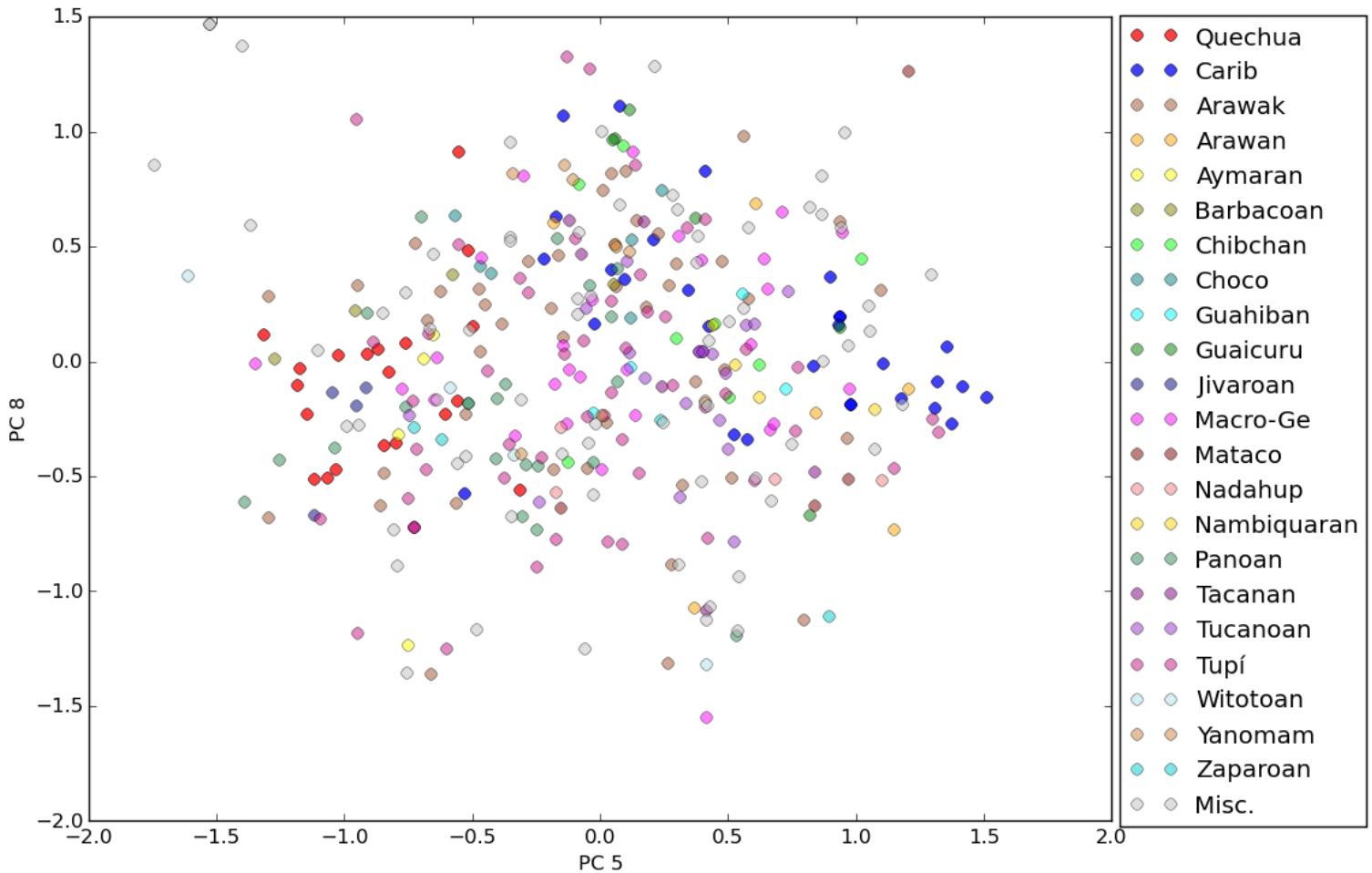
# PC5: genetic signal

- PC5 shows a west/east divide illustrated by the Quechua ($p<1.0E-9$) and Carib ($p<1.0E-4$) families



# PC5: genetic signal

- PC5 shows a west/east divide illustrated by the Quechua ($p<1.0E-9$) and Carib ($p<1.0E-4$) families



# PC5: summary

- Positive segments: mid vowels
- Negative segments: palatals, affricates, ɨ
- There is a general divide, both genetic and areal, between a negative signal in the west and a positive signal in the east
  - The western negative region corresponds to the Central-Northern Andean and Circum-Andean sub-region identified by Michael et al. (2014)
- Overall, the signal is becoming weak by this point

# Interim Summary: Areal results

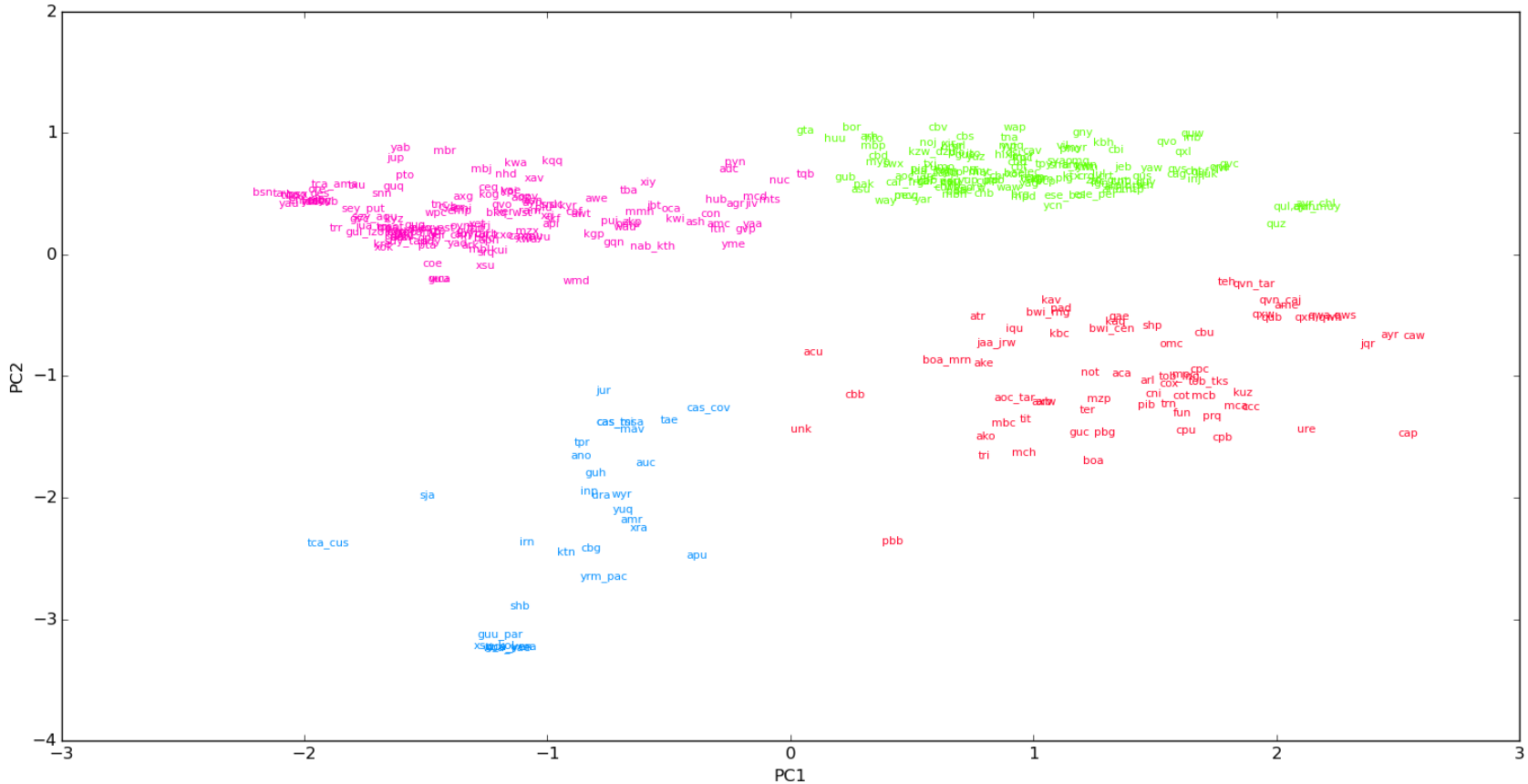
- Examination of geographical distribution of extremal values of PCs 1–5 have identified known phonological areas in South America:
  1. Andean and Circum-Andean phonological area
     - 1.1 Southern Andean and Circum-Andean phonological sub-area
     - 1.2 Central-Northern Andean and Circum-Andean phonological sub-area
  2. Northwest Amazonian phonological area
- Whereas previous identification of these areas have relied on methods that require human insight and intuition (Naive Bayes Classifier approaches; Michael et al. (2014)), in the PCA approach can these results emerge more directly from the data

# Identifying phonological 'types'

- Since the PCs identify the major dimensions of phonological variation in South America, we can develop a 'continental typology' of phonological inventories
- Major phonological types can be identified by examining how inventories cluster with respect to PC1 and PC2
  - Hierarchical clustering using a Euclidean distance measure and Ward's clustering criterion
- Sampling languages in each cluster allows us to identify their major features

# Phonological types in South America

- What major types can we extrapolate from PCA?
- What are the clusters and what are their general profiles?



# Quadrant 1: +PC1, +PC2

- Features: large consonant inventories (laterals, affricates, voiced stops) and small vowel inventories
- Example language: Salasaca Quechua (Quechua)

| **Consonants** | Bilabial | Alveolar | Post-alveolar | Palatal | Velar |
|---|---|---|---|---|---|
| Aspirated stop | pʰ | tʰ | | | kʰ |
| Plain/voiced stop | p b | t d | | | k g |
| Affricate | | ts | tʃ | | |
| Fricative | | s z | ʃ ʒ | | x |
| Nasal | m | n | | ɲ | |
| Approximant | | | | j | w |
| Tap, flap | | ɾ | | | |
| Lateral | | l | | | |

| **Vowels** | Front | Central | Back |
|---|---|---|---|
| High | i | | u |
| Low | | a | |

# Quadrant 2: +PC1, -PC2

- Features: large consonant inventories (laterals, affricates), moderate vowel inventories (long vowels)
- Example language: Chamicuro (Arawak)

| **Consonants** | Bilabial | Alveolar | Post-alveolar | Retroflex | Palatal | Velar | Glottal |
|---|---|---|---|---|---|---|---|
| Stop | p | t | | | | k | ʔ |
| Affricate | | ts | tʃ | tʂ | | | |
| Fricative | | s | ʃ | ʂ | | | h |
| Nasal | m | n | | | ɲ | | |
| Approximant | | | | | j | w | |
| Tap, flap | | ɾ | | | | | |
| Lateral | | l | | | ʎ | | |

| **Vowels** | Front | Central | Back |
|---|---|---|---|
| High | i iː | | u uː |
| Mid | e eː | | o oː |
| Low | | a aː | |

# Quadrant 3: -PC1, -PC2

- Features: small consonant inventories, large vowel inventories (nasal and long vowels)
- Example language: Karitiâna (Tupí)

| **Consonants** | Bilabial | Alveolar | Palatal | Velar | Glottal |
|---|---|---|---|---|---|
| Stop | p | t | | k | |
| Fricative | | s | | | h |
| Nasal | m | n | ɲ | ŋ | |
| Approximant | | | | w | |
| Tap, flap | | ɾ | | | |

| **Vowels** | Front | Central | Back |
|---|---|---|---|
| High | i ĩ i: ĩ: | ɨ ɨ̃ ɨ: ɨ̃: | |
| Mid | e ẽ e: ẽ: | | o õ o: õ: |
| Low | | a ã a: ã: | |

# Quadrant 4: -PC1, +PC2

- Features: moderate consonant inventories (voiced stops), moderate vowel inventories (nasal vowels)
- Example language: Siona (Tucanoan)

| **Consonants** | Bilabial | Alveolar | Post-alveolar | Palatal | Velar | Labio-velar | Glottal |
|---|---|---|---|---|---|---|---|
| Stop/affricate | p b | t d | tʃ | | k g | $k^w$ $g^w$ | ʔ |
| Fricative | | s z | | | | $h^w$ | h |
| Nasal | m | n | | | | | |
| Approximant | | | | j | | w | |

| **Vowels** | Front | Central | Back |
|---|---|---|---|
| High | i ĩ | ɨ ɨ̃ | u ũ |
| Mid | e ẽ | | o õ |
| Low | | a ã | |

# Conclusion

- In South America, whether languages make a large number of contrasts in their vowels is very significant in 'typing' languages
  - Nasal vs. oral is one of the most significant dimensions of variation
  - Length contrasts are also an important parameter of differentiation
- PCA is successful in producing this continental typology as well as in identifying important linguistic areas such as the Andean and Circum-Andean region and Northwest Amazonia
- This work provides a starting point for more quantitatively rigorous analyses of areality

Aikhenvald, Alexandra Y. 2002. *Language contact in Amazonia*. Oxford: Oxford University Press.

Michael, Lev, Will Chang, and Tammy Stark. 2014. Exploring phonological areality in the cirum-Andean region using a naive Bayes classifier. *Language Dynamics and Change* 4:27–86.

Michael, Lev, Tammy Stark, Emily Clem, and Will Chang. 2016. *South American phonological inventory database*. Berkeley: University of California.

O'Connor, Loretta, and Pieter Muysken, ed. 2014. *The native languages of South America: Origins, development, typology*. Cambridge: Cambridge University Press.

Reesink, Ger, Ruth Singer, and Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biol* 7.

Wichmann, Søren, and Jeff Good, ed. 2014. *Quantifying language dynamics: On the cutting edge of areal and phylogenetic linguistics*. Leiden: Brill.